Demystifying Video Reasoning

Ruisi Wang[1], Zhongang Cai[1,✉], Fanyi Pu[1,2], Junxiang Xu[1], Wanqi Yin[1], Maijunxian Wang[3], Ran Ji[4], Chenyang Gu[1], Bo Li[2], Ziqi Huang[2], Hokin Deng[5], Dahua Lin[1], Ziwei Liu[2], Lei Yang[1]

[1] SenseTime Research
[2] Nanyang Technological University [3] University of California, Berkeley
[4] University of California, San Diego [5] Carnegie Mellon University
{wangruisi1,caizhongang}@sensetime.com

Abstract. Recent advances in video generation have revealed an unexpected capability: diffusion-based video models can perform reasoning through Chain-of-Frames (CoF), suggesting that reasoning unfolds sequentially across video frames. In this work, we revisit this assumption and uncover a different mechanism. Through systematic analysis, we show that video reasoning primarily develops along the diffusion denoising steps instead, where early steps exhibit multiple hypotheses that gradually converge. We term this mechanism **Chain-of-Steps (CoS)**, which is validated through qualitative analysis and probing tests. Our investigation reveals several intriguing emergent behaviors: 1) Models demonstrate a form of persistent **working memory** that supports tasks requiring consistent reference, such as object permanence. 2) They can **self-correct** intermediate mistakes during generation instead of committing to incorrect trajectories. 3) Early steps function differently from later steps, exemplified by a **perception before action** phenomenon. Moreover, analysis of Diffusion Transformer layers shows that middle layers conduct key reasoning procedures. Motivated by these insights, we propose **Training-Free Ensemble (TFE)**, a simple strategy that integrates reasoning paths by merging latents from identical models with different random seeds at inference time. This approach encourages the exploration of diverse reasoning trajectories and improves reasoning performance. Together, our findings provide the first systematic dissection of the mechanisms underlying video reasoning and offer practical insights for developing more capable video reasoning models.

Keywords: Video Reasoning · Diffusion Models · Emergent Intelligence

1 Introduction

Video generation models have transformed the landscape of movie, gaming, and entertainment industries. However, most research has focused primarily on their

✉ Corresponding Author;
Homepage: https://www.wruisi.com/demystifying_video_reasoning.

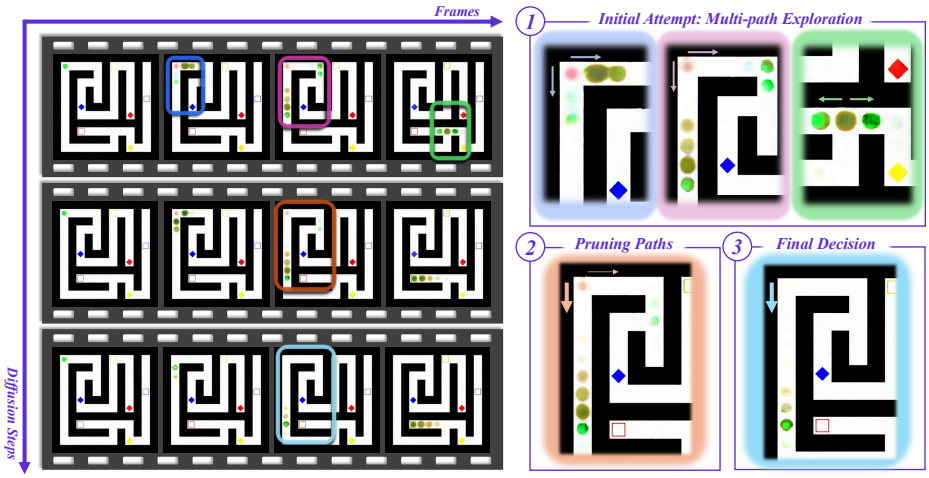

Fig. 1: Chain-of-Steps. We discover that video reasoning occurs along the diffusion steps with surprising emergent behaviors such as making multiple possible moves (*e.g.*, paths) simultaneously at early steps, gradually pruning suboptimal choices during middle steps, and reaching a final decision at the late steps. This maze-solving example asks the model to start from the green circle in the top-left corner and find the red rectangle. Key regions of interest are color-coded and enlarged on the right.

ability to produce high-fidelity, realistic, and visually appealing videos. Recent advances have revealed an unexpected phenomenon: diffusion-based video models exhibit non-trivial reasoning capabilities in spatiotemporally consistent visual environments [65]. Prior work attributes this behavior to a Chain-of-Frames (CoF) mechanism, suggesting that reasoning unfolds sequentially across video frames. Despite this intriguing discovery, the underlying mechanisms of video reasoning remain largely unexplored. With the recent release of large-scale video reasoning datasets and open-source foundation models [61], we can now systematically investigate this capability. Leveraging these resources, we conduct the first comprehensive dissection of video reasoning and uncover a fundamentally different mechanism: **reasoning in diffusion-based video models primarily emerges along the denoising process** rather than across frames.

Our key discovery challenges the prevailing Chain-of-Frames (CoF) hypothesis [65, 69], which assumes that video reasoning unfolds sequentially across frames. Instead, we find that reasoning does not primarily operate along the temporal dimension. Rather, it emerges along the diffusion denoising steps, progressing throughout generation. We term this mechanism **Chain-of-Steps (CoS)**. This finding suggests a fundamentally different view of how diffusion-based video models reason. Due to bidirectional attention over the entire sequence, reasoning is performed across all frames simultaneously at each denoising step, with intermediate hypotheses progressively refined as the process unfolds. Qualitative analysis reveals intriguing dynamics. In early denoising steps, the model often entertains multiple possibilities, exploring alternative trajectories before progressively committing to one. Importantly, we view reasoning not as the mere

coexistence of multiple possible outcomes, but as a directed process of selecting and refining a solution from among candidates. Moreover, noise perturbation analysis shows that disruptions at specific denoising steps significantly degrade performance, whereas frame-wise perturbations have a much weaker impact. Further information propagation analysis identifies that conclusions primarily solidify during the middle diffusion steps.

Furthermore, we uncover several surprising emergent behaviors in video reasoning models that are strikingly similar to those observed in early studies of Large Language Models (LLMs). First, these models exhibit a form of **working memory** that is crucial for tasks requiring persistent references (*e.g.*, object permanence). Second, we observe that video models can **self-correct** errors during the CoS reasoning process, rather than committing to incorrect trajectories. Third, video models exhibit a **"perception before action"** behavior, where early diffusion steps prioritize localizing target objects before subsequent steps perform more complex reasoning and manipulation.

We further conduct a fine-grained analysis of the Diffusion Transformer by examining token representations within a single diffusion step. This reveals self-evolved, diverse, task-agnostic functional layers throughout the network. Within one step, early layers focus on dense perceptual understanding (*e.g.*, separating foreground from background and identifying basic geometric structures), while a set of critical middle layers performs most of the reasoning. The final layers then consolidate the latent representation to produce the next step video state.

Motivated by these insights, we present Training-Free Ensemble (TFE), a simple method as a proof-of-concept for improving video reasoning models. Given that the model inherently explores multiple reasoning paths during the diffusion process, we propose an inference-time ensemble strategy that merges latents produced by three identical models with different random seeds. This encourages the preservation of a richer set of candidate reasoning trajectories during generation. As a result, the model explores more diverse reasoning paths and is more likely to converge to the correct solution, illustrating how our findings can inform more effective video reasoning systems.

In summary, we investigate the underlying mechanisms of video reasoning in diffusion models and identify Chain-of-Steps (CoS), a reasoning process that unfolds along the denoising trajectory. We further uncover several emergent reasoning behaviors that arise in these models. Building on these insights, we demonstrate how such mechanisms can be exploited through a simple training-free strategy for reasoning path ensembling. We believe our findings provide a foundation for understanding and advancing video reasoning, positioning it as a promising next-generation substrate for machine intelligence.

2 Related Works

2.1 Reasoning in Language and Multimodal Models

Recent studies show that large language models (LLMs) exhibit remarkable reasoning capabilities. Early work identifies emergent behaviors that arise as

models scale in size and data [63], and demonstrates that Chain-of-Thought (CoT) prompting, which elicits intermediate reasoning steps, significantly improves performance [64]. Subsequent work explores mechanisms such as self-reflection, correction, and action [22, 42, 72, 75]. Coconut further suggests that reasoning can also occur implicitly within latent representations [19]. Meanwhile, research has increasingly explored extending reasoning beyond language into multimodal settings. Early progress in vision-language models (VLMs) enables reasoning over images in addition to text [1, 2, 6, 33, 38], whereas recent work has studied unified architectures that jointly model language and vision [4, 8, 14, 34, 45, 49, 57, 59, 67, 68, 83, 84]. These architectures empower reasoning for generation [10, 17, 28, 35, 53, 70, 82], enable reasoning with generation through visual CoT [5, 7, 13, 24, 26, 32, 48, 54, 71], and extend to embodied scenarios [77–79]. Together, these findings suggest that reasoning over multimodal signals opens up avenues for advanced reasoning capabilities. However, these efforts remain limited to discrete text and static images, making it challenging to leverage spatiotemporally consistent priors. Our work aims to investigate video as the next substrate for reasoning in intelligent systems.

2.2 Video Generation Models

Video generation has advanced rapidly with the development of diffusion models [21, 50] and high-fidelity variational autoencoders (VAEs) [9, 27, 80]. While early approaches focus primarily on generating short clips, the emergence of Diffusion Transformers (DiTs) [46] has enabled effective scaling of data and model size. As a result, recent video generators [11, 15, 29, 31, 44, 51, 60, 62] achieve impressive visual fidelity. Despite these advances, major challenges remain in physical plausibility [76, 81], commonsense knowledge [81], and spatiotemporal reasoning [40, 61, 65]. Consequently, recent research has begun shifting toward investigating the reasoning capabilities of video generation models. One line of work leverages the reasoning abilities of multimodal LLMs to guide video synthesis. For example, VChain [24] and MetaCanvas [36] incorporate external reasoning modules into pre-trained generators, while Omni-Video [56] uses symbolic reasoning from LLMs to guide generation. More recently, several studies ask whether video generators themselves can perform reasoning without external supervision, treating them as zero-shot learners operating in spatiotemporal environments [20, 58, 65]. However, the mechanisms underlying this capability remain unexplored. Our work addresses this gap by investigating the internal reasoning processes of diffusion-based video models.

2.3 Similarities to Biological Brains

The diffusion model may exhibit behavior analogous to planning processes in biological brains. For instance, when a rat is deciding how to navigate toward a food reward, studies have shown that the hippocampus sequentially replays multiple candidate trajectories during a deliberation period before movement begins [47]. Recent work suggests that human cognition may rely on analogous

mechanisms of internal simulation and prospective planning to support conceptual reasoning, inference, and decision-making [3, 43].

3 Chain-of-Steps: Reasoning along Diffusion Steps

While prior work [65] hypothesizes a *Chain-of-Frames* (CoF) mechanism in which reasoning in video models unfolds frame by frame, generated frames appear to exhibit a "causal" property where later frames gradually build conclusions conditioned on earlier frames. However, our analysis of the underlying video reasoning mechanism reveals evidence to the contrary. First, we empirically analyze a wide range of reasoning tasks and find that the core logical reasoning in video generation models occurs across the diffusion denoising steps (Sec. 3.1). Diffusion steps do more than merely refine visual texture; instead, they explore multiple possibilities, evaluate their plausibility, and gradually converge to the correct outcome through the denoising process. Second, we introduce noise perturbations to disrupt information flow at both the frame and step levels (Sec. 3.2). Our findings reaffirm that CoS, rather than CoF, more accurately characterizes the reasoning mechanism in video models. To ensure generality, we analyze multiple base models (*i.e.*, LTX2.3 [18], Wan2.1 [60], and Wan2.2 [60]) and their reasoning-enhanced variants (with LoRA tuned on VBVR [61], indicated by the **VBVR-** prefix) that are significantly better at instruction-following and precise manipulation. Unless otherwise stated, visualizations shown in the main paper are based on VBVR-Wan2.2, which is the strongest video reasoning model based on Wan2.2-I2V-A14B [60], while results for more models are provided in Sec. E.

3.1 Diffusion Steps as the Primary Axis of Reasoning

To understand how semantic decisions emerge during denoising, we seek to visualize the model's intermediate trajectories throughout the diffusion process. However, this is challenging because diffusion models do not explicitly expose interpretable reasoning states. Directly decoding the intermediate latent x_s is noisy and uninformative, as flow matching [37] constructs x_s as a linear interpolation between semantic content and stochastic noise,

$$x_s = (1 - s)x_0 + sx_1 \tag{1}$$

where x_0 is the clean latent and $x_1 \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ is noise. To obtain a more interpretable view, we instead compute the estimated clean latent, using the predicted velocity field $v_\theta(x_s, s, c)$ and the corresponding noise scale σ_s.

$$\hat{x}_0 = x_s - \sigma_s \cdot v_\theta(x_s, s, c) \tag{2}$$

Importantly, $\hat{x}_0$ is not a latent encountered during inference, but rather the model's instantaneous prediction of the final denoised sample given all information available at diffusion step s. Decoding $\hat{x}_0$ therefore provides an interpretable

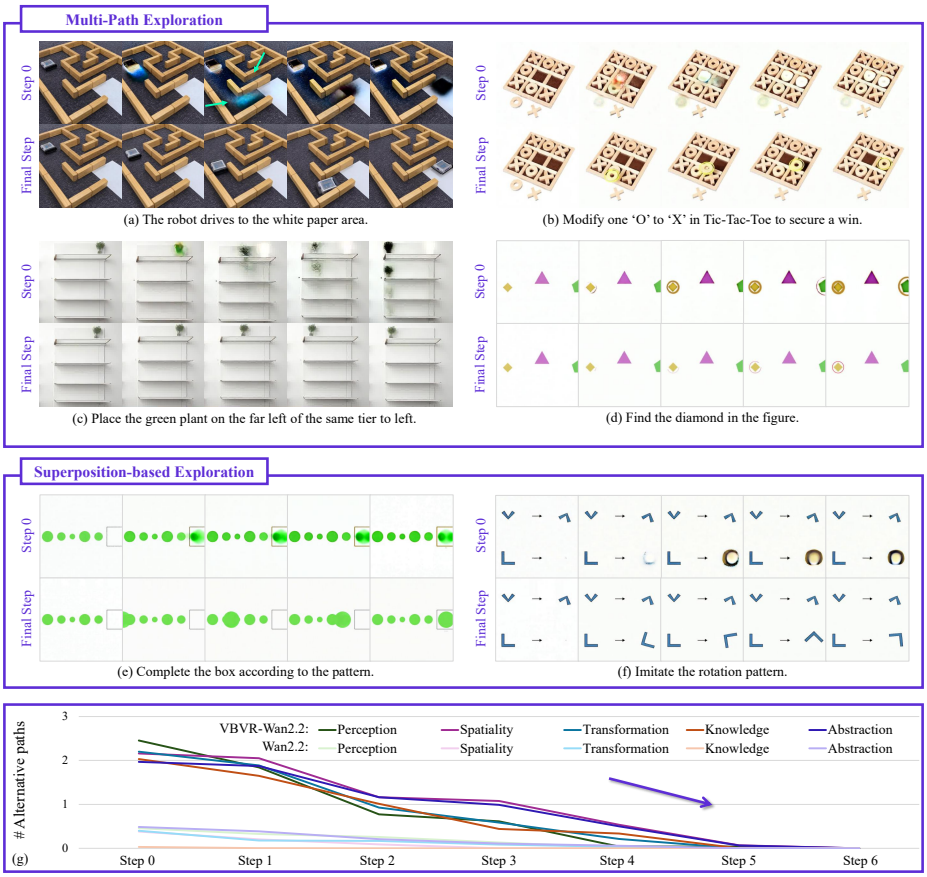

Fig. 2: Chain-of-Steps elicits reasoning along the diffusion process. We observe that video reasoning models explore multiple possible solutions simultaneously in the early denoising steps before converging to a final outcome in later steps. Specifically, we observe: (a) two potential routes (cyan arrows highlight the "imaginary traces") for the robot; (b) two possible placements of the "O" piece; (c) multiple candidate end positions for the plant; (d) simultaneous selection of two diamonds; (e) large and small circles overlapping with each other; and (f) all possible rotations of the L-shaped object superimposed. (g) Average number of alternative intermediate candidates observed at each diffusion step.

snapshot of the model's current hypothesis, enabling us to trace how semantic decisions evolve and analyze the model's intermediate reasoning dynamics.

Visualization examples are drawn from video reasoning benchmarks such as VBVR [61] and general video generation benchmarks such as VBench [23, 25]. Analogous to LLMs that exhibit reasoning behaviors along chain-of-thought, where the model gradually reaches its conclusion, to our surprise, we discover a similar scheme in video reasoning models along diffusion denoising steps. This is exemplified in Fig. 1, for complex navigational tasks such as maze-solving, early

decoded predictions $\hat{x}_0$ appear as a probabilistic cloud in which several plausible paths are spawned and explored in parallel. Over subsequent steps, suboptimal trajectories gradually get suppressed, converging towards the final solution. By analyzing intermediate predictions at each step, we move beyond the temporal "Chain-of-Frames" (CoF) perspective and identify two distinct modes of Stepwise Reasoning: Multi-path Exploration and Superposition-based Exploration.

Multi-Path Exploration. In high-complexity logical tasks, the diffusion process resembles a Breadth-First Search (BFS) or a multi-choice elimination procedure, where the model explores a tree of possible solutions and gradually prunes incorrect branches. It is worth noting that this behavior is reminiscent of the parallel reasoning trajectories explicitly studied in the LLM community (*e.g.*, Tree of Thoughts [74]). However, video generation models naturally explore multiple solution paths in parallel during the diffusion process, inherently performing a similar form of structured search within their latent space. In some tasks involving object movements, the model explicitly visualizes this exploration process through multiple motion trajectories. In other tasks where the model must select an action from a discrete set of alternatives, we observe that the model initially considers several actions simultaneously and progressively discards candidates as the denoising process proceeds, until only a single valid outcome remains.

- *Fig. 2(a) Robot Navigation.* The intermediate steps show the robot simultaneously exploring both the upper and lower routes through the maze. As the diffusion process proceeds, the trajectory corresponding to the lower path becomes increasingly dominant, while the alternative route gradually disappears, indicating that the model chooses the final path.
- *Fig. 2(b) Tic-Tac-Toe.* During the early reasoning stage, the model simultaneously highlights multiple candidate cells for a winning move.
- *Fig. 2(c) Object Movement.* In this example, it is clearly observable that at the early stage, the model proposes four potential trajectories corresponding to the four layers on the left side of the shelf. As the denoising steps continue, these alternatives gradually collapse toward placing the plant on the first layer, producing a clear and consistent motion path.
- *Fig. 2(d) Diamond Detection.* The model initially marks two candidate shapes that might satisfy the query. Through iterative refinement, the incorrect candidate fades; only the correct diamond remains circled in the end.

Superposition-based Exploration. Another distinctive mode observable along the diffusion trajectory is superposition-based exploration, where the model temporarily represents multiple mutually exclusive logical states simultaneously. Instead of committing early to a single configuration, the model maintains overlapping hypotheses that gradually resolve as noise is removed. This phenomenon is particularly evident in tasks involving object reordering and spatial alignment, where alternative outcomes often occupy similar spatial regions and overlapping pixel regions, making superposed hypotheses directly visible.

- *Fig. 2(e) Size Pattern Completion.* The size-pattern follows a repeating "large-medium–small" pattern. When predicting the next element, the model initially generates overlapping circles of different sizes, representing competing hypotheses about the correct continuation of the sequence.
- *Fig. 2(f) Object Rotation.* Instead of a discrete rotation from one angle to another, the model produces a blurred superposition of multiple orientations.

Discussion of Reasoning Patterns. We examine whether the observed *multi-path exploration* and *superposition-based exploration* reflect underlying reasoning patterns in video models. **Prevalence.** To assess the prevalence of these phenomena, we manually inspect 200 examples drawn from VBVR-Bench and real-world videos. We find that 72% exhibit either *multi-path exploration* or *superposition-based exploration* during denoising. The remaining examples typically involve simpler tasks whose trajectories converge directly to a solution. **Validity.** We further argue that these phenomena cannot be fully attributed to generic diffusion artifacts or uncertainty. Conventional artifacts, such as blur and ghosting, generally lack clear semantic structure. In contrast, the early-stage candidates we observe are semantically coherent and task-consistent. Moreover, as shown in Fig. 2, we quantify the degree of intermediate exploration using *candidate multiplicity*. Specifically, at each denoising step, we count the number of semantically distinct and task-consistent candidate solutions visible in the intermediate output. We find that this behavior is substantially more pronounced in the reasoning-enhanced VBVR-Wan2.2 model than in general video-generation baselines. Together, these findings suggest that the observed patterns reflect structured exploration beyond ordinary diffusion ambiguity.

3.2 Noise Perturbation and Information Flow

Our hypothesis is further validated through targeted noise injection experiments. We compare two settings to isolate where the core reasoning process occurs: 1) "Noise at Step": $x_{s,\forall f} \leftarrow \mathcal{N}(0, \mathbf{I})$. That is, disruptive Gaussian noise is injected into all frames at a specific diffusion step. 2) "Noise at Frame": $x_{\forall s,f} \leftarrow \mathcal{N}(0, \mathbf{I})$. That is, Gaussian noise is injected into a specific frame across all diffusion steps. The two settings are illustrated in Fig. 3(a).

In Fig. 3(b), we evaluate model performance under these two noise injection schemes. Compared to the baseline without noise, the "Noise at Step" setting causes the final score to collapse from 0.685 to below 0.3, indicating that the reasoning trajectory is highly sensitive to disruptions along the diffusion steps. Noise injected at a particular diffusion step therefore leads to a significant interruption of the model's reasoning process.

In contrast, under "Noise at Frame" injection, the model demonstrates more robustness with a much smaller performance drop. This behavior can be explained by the architecture of diffusion transformers: each denoising step has full observation of the preceding latent sequence through bidirectional attention, allowing the model to refine the entire video latent jointly. Consequently,

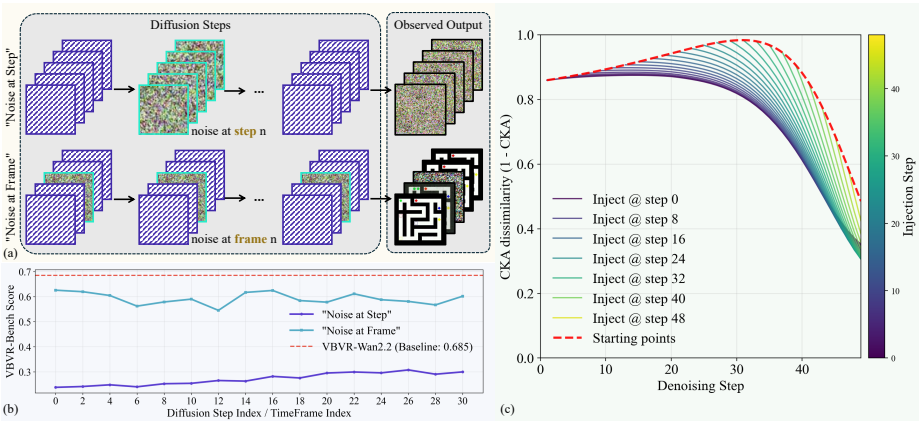

Fig. 3: Noise perturbation and information flow. (a) Illustration of noise injection schemes; "Noise at Step" suffers more significant corruption than "Noise at Frame". (b) Performance drop with the two noise injection schemes. X-axis is the injection index (either diffusion step or frame). (c) Information flow across denoising steps (CKA dissimilarity: 1.0 indicates complete corruption, 0.0 indicates no effect).

corrupted frames can be recovered by leveraging the uncorrupted information from neighboring frames during subsequent denoising steps.

In Fig. 3(c), we further analyze information propagation by measuring divergence after injecting noise at step s_t. We visualize CKA dissimilarity [30], where 1.0 indicates complete corruption and 0.0 indicates no effect. The results show that perturbations introduced in early diffusion steps propagate throughout the entire trajectory, fundamentally altering the final reasoning outcome.

Moreover, the red dotted line highlights step-wise sensitivity to disruptive noise, which gradually increases and peaks around steps 20–30. This observation aligns with our qualitative analysis. Although steps 20–30 are not the earliest stages where we first observe reasoning phenomena, by this point the model has already pruned its reasoning trajectory toward the final conclusion. Consequently, perturbations at these steps have a large impact, as they can disrupt a reasoning process that is nearly finalized. Later steps, in contrast, appear less critical for the model's reasoning capability.

4 Emergent Reasoning Behaviors

Similar to the emergent reasoning behaviors observed in Large Language Models (LLMs) [22,42,64,72,75], we identify three surprising properties that are critical to effective video reasoning: *working memory* (Sec. 4.1), which retains essential information throughout the reasoning process; *self-correction and enhancement* (Sec. 4.2), which enables the model to revise intermediate hypotheses or refine previously generated answers, gradually adjusting toward the optimal solution even when it is not present initially; and *perception before action* (Sec. 4.3),

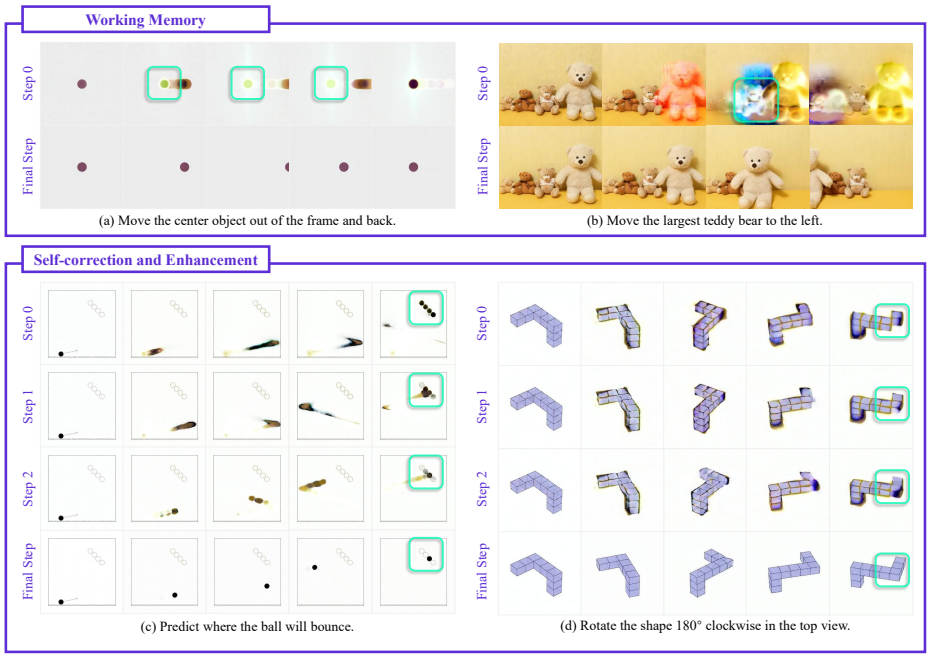

Fig. 4: Emergent reasoning behaviors: memory and self-correction. (a) The center point is retained to guide the return motion. (b) The contour of the occluded small teddy bear is preserved, enabling the model to address object permanence. (c) The trajectory of the ball gradually extends and becomes complete. (d) The missing cube only appears in the later diffusion steps. Cyan boxes are added for illustration; they are not part of the generated video.

indicates that the model first establishes a grounded understanding of scene entities and spatial layout before reasoning about actions and interactions.

4.1 Working Memory

Reasoning requires the maintenance of "working memory" or a state. The demonstrations show that the diffusion process naturally establishes persistent anchors that preserve critical information across generation steps.

- *Fig. 4(a) Object Reappearance.* The model consistently preserves the object's initial position throughout the diffusion steps, enabling the circle to return to its original location and remain consistent with the initial condition.
- *Fig. 4(b) Teddy Bear Relocation.* During the movement task, the largest teddy bear temporarily blocks the small teddy bear on the left. Despite this occlusion, the early diffusion steps retain the state of the small bear to ensure consistent generation in the whole video.

4.2 Self-correction and Enhancement

During the diffusion process, we observe several stochastic "aha moments," where the model initially selects an incorrect option but later revises its reasoning after

a few diffusion steps, exploring an alternative strategy. These behaviors are functionally analogous to the internal backtracking and "slow thinking" discussed in long-thinking Large Language Models (LLMs) [72]. Importantly, such transitions are not limited to correcting mistakes. The model may also refine an initially incomplete answer into a logically richer and more comprehensive one, reflecting a form of latent self-improvement rather than simple error repair.

In contrast to the "Chain-of-Frames" theory, which would require such corrections to happen sequentially across time, these reversals take place globally across all frames simultaneously within a single diffusion step. This further provides strong evidence that the video generation model prioritizes global logical integrity over local, sequential frame-wise updates.

- *Fig. 4(c) Hit Target After Bounce.* Initially, the ball's trajectory is incomplete and ambiguous. As diffusion progresses, the model gradually completes the trajectory, making it increasingly clear, and the outcome converges from four candidate points to a single correct point.
- *Fig. 4(d) 3D Shape Rotation.* At the first diffusion step, the rotated cubes are generated with incorrect quantities and arrangements. After several diffusion steps, the model gradually corrects both the number and the spatial configuration, producing a coherent and accurate final result.

4.3 Perception before Action

We observe a phenomenon suggesting that the diffusion trajectory first addresses the "what" and "where" of a scene before determining the "how" and "why" of its thinking progression. This reflects a "Perception before Action" transition, where the model progresses from static grounding to dynamic reasoning. As illustrated in Fig. 5, the initial diffusion step primarily identifies the foreground entity specified in the prompt (*e.g.*, the car or the door), without exhibiting explicit motion planning or relational transformation. Dynamic structure emerges only in later steps, as the model moves beyond scene grounding and starts coordinating object motion and inter-object interactions.

5 Layer-wise Mechanistic Analysis

Inspired by the discovery of vision function layers in vision-language models [52], we investigate how diffusion transformers process visual information during video reasoning by analyzing the internal representations across transformer layers. Rather than focusing solely on generated outputs, we examine how hidden states evolve within the DiT backbone and how different layers contribute to semantic grounding and reasoning behaviors. Specifically, we study the model from two complementary perspectives. First, we visualize token-level activations across layers to analyze how representation energy distribute over spatial-temporal regions. Second, we perform layer-wise latent swapping experiments to causally

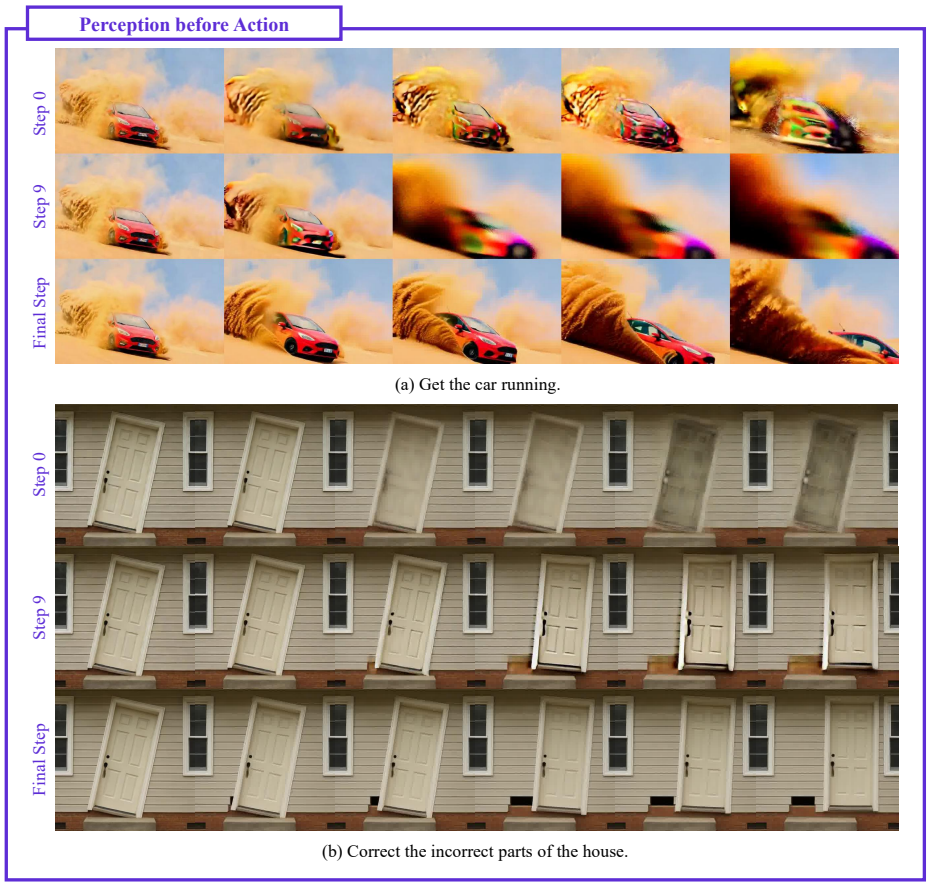

Fig. 5: Emergent reasoning behavior: understanding before reasoning. (a) Early diffusion steps identify the car as the object of interest with no motion, while later steps introduce action and simulate physical interactions. (b) Early steps recognize the door as the target object with no motion, and later steps manipulate it.

evaluate how intermediate representations influence the final reasoning outcome. Together, these analyses provide a fine-grained view of how information is organized and progressively transformed inside the model.

5.1 Layer-wise Token-Level Visualization

To further investigate this transition, we analyze the internal activations of DiT blocks. During each diffusion step, forward hooks are registered on all transformer blocks of Wan2.2-I2V-A14B, and hidden states from the first forward pass are extracted to isolate the model's primary reasoning trajectory. The captured features are represented as a token sequence $feat \in \mathbb{R}^{B\times N\times D}$, where N represents the total number of tokens and $D = 5120$ is the embedding dimension. To restore the visual context, we utilize the grid dimensions (f, h, w) captured

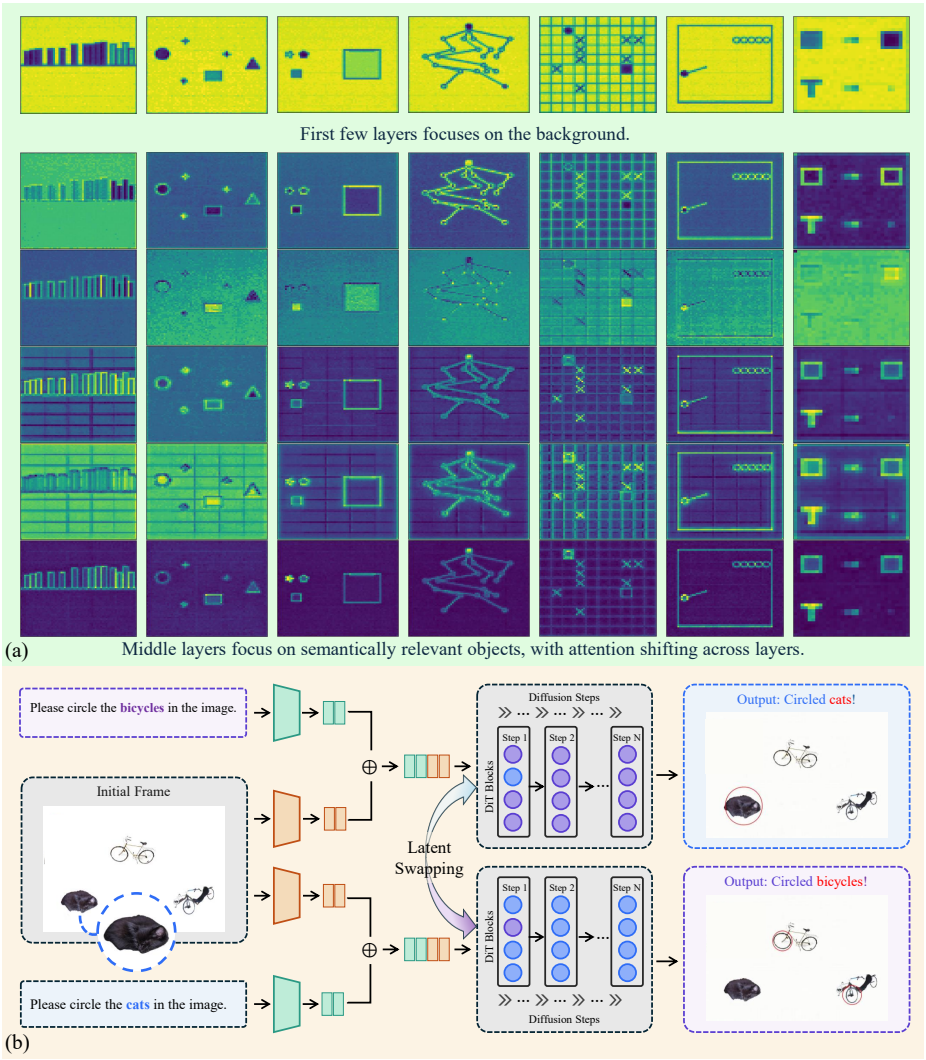

Fig. 6: Layer specialization. (a) Layer-wise activation visualization shows that early layers of the video reasoning DiT tend to focus on background structures, whereas later layers carry out reasoning-related computations. (b) Layer-wise latent swapping reveals that certain middle layers (*e.g.*, Layer 21 in this case) contain critical reasoning information that strongly influences the final outcome.

from the model's `patch_embedding` layer to reshape the features into a 5D tensor of shape (B, f, h, w, D). Activation strength is then quantified by the channel-wise L_2 norm is computed along the channel dimension for each token, producing token-level energy maps that are visualized across layers and video frames. In Fig. 6(a), rows correspond to selected DiT layers (e.g., $L0, L10, L20 \ldots L39$) and columns represent sequential video frames. Each cell in the grid displays a

heatmap of token energies, enabling visualization of how the model's attention shifts throughout the network.

Within a single diffusion step, the earliest layers (Layers 0–9) primarily attend to global structures and background context. As computation proceeds through the layers of the same step, attention progressively shifts toward foreground entities and those specified in the prompt. From around Layer 9 onward, activations become increasingly concentrated on semantically relevant objects, accompanied by higher channel variance in localized regions corresponding to target entities. Notably, reasoning-related features also begin to emerge at this stage, with activations correlating with object motion and interactions. For example, in Column 4 of Fig. 6(a), the activation pattern evolves from the background to all nodes and paths, then narrows to the nodes themselves before revisiting the complete graph, Finally, the activation becomes concentrated on the starting node. This within-step progression is consistently observed across diffusion steps, indicating a recurrent hierarchy from global context to object-centric reasoning.

5.2 Layer-wise Latent Swapping Experiment

As illustrated in Fig. 6(b), we perform a layer-wise latent swapping experiment to understand which transformer layers are responsible for determining the final grounding result. The experiment is performed on an object grounding task at the first diffusion step using a controlled setup of blank background and two object configurations (O_A, O_B). Two forward passes are first obtained using O_A and O_B, producing latent representations $U_A^{(l)}$ and $U_B^{(l)}$ at each transformer layer. Then, hybrid forward passes for each layer k are constructed by replacing only the representation at layer k with that from the alternative configuration while keeping all other layers unchanged: $\tilde{U}^{(k)} \leftarrow U_{alt}^{(k)}$, and $U^{(l \neq k)} = U_{orig}^{(l)}$.

Experiments are conducted on ten additional object categories across five seeds. Remarkably, swapping representations at one middle layer (around 15 to 35) is sufficient to reverse the grounding result, causing the predicted target object to flip. This suggests that middle-to-late layers contain semantically decisive information for object grounding. A broader layer-wise analysis of flip rates is provided in Appendix Sec. B.1.

6 Training-Free Ensemble

To further validate our key observations (*e.g.*, multi-path reasoning during the early diffusion steps in Sec. 3.1 and Sec. 4.3, and reasoning-active features emerge in the mid-layers in Sec. 5), we design a simple proof-of-concept method, which we call Training-free Ensemble (TFE). Even simpler than merging models within the same optimization basin [66], we implement a multi-seed ensemble of the same model at the latent level during the *early* diffusion steps, with the hypothesis that the reasoning manifold (the internal latent space) contains a shared probabilistic bias toward the correct outcome. Specifically, we execute three independent forward passes using different initial noise seeds. During the first

Table 1: Benchmarking results on VBVR-Bench. Overall In-Domain (ID) and Out-of-Domain (OOD) scores are reported alongside category-wise performance. Higher is better. **Bold**: best in group; underline: second best.

			In-Domain by Category					Out-of-Domain by Category					
Models	**Overall**	**Avg.**	**Abst.**	**Know.**	**Perc.**	**Spat.**	**Trans.**	**Avg.**	**Abst.**	**Know.**	**Perc.**	**Spat.**	**Trans.**
Human	0.974	0.960	0.919	0.956	1.00	0.95	1.00	0.988	1.00	1.00	0.990	1.00	0.970
Open-source Video Models													
CogVideoX1.5-5B-I2V [73]	0.273	0.283	0.241	0.328	0.257	0.328	0.305	0.262	0.281	0.235	0.250	**0.254**	0.282
HunyuanVideo-I2V [29]	0.273	0.280	0.207	0.357	0.293	0.280	0.316	0.265	0.175	**0.369**	0.290	0.253	0.250
Wan2.2-I2V-A14B [60]	**0.371**	**0.412**	**0.430**	**0.382**	**0.415**	**0.404**	**0.419**	**0.329**	**0.405**	0.308	**0.343**	0.236	0.307
LTX-2 [18]	0.313	0.329	0.316	0.362	0.326	0.340	0.306	0.297	0.244	0.337	0.317	0.231	**0.311**
Proprietary Video Models													
Runway Gen-4 Turbo [51]	0.403	0.392	0.396	0.409	0.429	0.341	0.363	0.414	0.515	0.429	0.419	0.327	0.373
Sora 2 [44]	**0.546**	**0.569**	0.602	0.477	**0.581**	**0.572**	**0.597**	**0.523**	0.546	**0.472**	**0.525**	**0.462**	**0.546**
Kling 2.6 [31]	0.369	0.408	0.465	0.323	0.375	0.347	0.519	0.330	0.528	0.135	0.272	0.356	0.359
Veo 3.1 [15]	0.480	0.531	**0.611**	**0.503**	0.520	0.444	0.510	0.429	**0.577**	0.277	0.420	0.441	0.404
Video Reasoning Models													
VBVR-Wan2.2 [61]	0.685	0.760	0.724	**0.750**	0.782	0.745	0.833	0.610	0.768	0.572	**0.547**	0.618	0.615
VBVR-Wan2.2 + Training-Free Ensemble	**0.716**	**0.780**	**0.760**	0.744	**0.809**	**0.749**	**0.858**	**0.650**	**0.803**	**0.705**	0.531	**0.639**	**0.716**

diffusion step ($s = 0$), we extract the hidden representations $U^{(l)}$ from the transformer backbone, and perform a spatial-temporal averaging of the latents across mid-layers: 20 to 29. By aggregating representations within this specific reasoning window, we effectively perform a latent-space ensemble resembling expert voting. This operation filters out seed-specific noise and biases the model's probability distribution toward a more stable and logically consistent latent state.

Table 2: TFE results on four benchmarks

Method	**VBVR-Bench**	**V-Reason Bench**	**MME-CoF**	**RISE-Video**
VBVR-Wan2.2	0.69	8.94	1.30	61.60
+ TFE	0.72	12.12	1.52	65.85
VBVR-Wan2.1	0.59	12.66	0.36	37.35
+ TFE	0.61	13.35	0.38	42.92
VBVR-LTX2.3	0.52	3.61	0.12	32.75
+ TFE	0.53	3.19	0.27	35.85

Tab. 1 shows that TFE improves VBVR-Wan2.2 across subcategories in VBVR-Bench [61]. Extending the evaluation to multiple video generation base models (LTX2.3, Wan2.1, and Wan2.2) across four benchmarks (VBVR-Bench, V-Reason Bench [41], MME-CoF [16], and RISE-Video [39]) shows similar trends in Tab. 2, where TFE consistently improves performance across nearly all model-benchmark pairs without any additional training. For instance, VBVR-Wan2.2 achieves relative improvements of 4.3% on VBVR-Bench, 35.6% on V-Reason Bench, 16.9% on MME-CoF, and 6.9% on RISE-Video. Similar improvements are observed for VBVR-Wan2.1 and VBVR-LTX2.3, demonstrating that the proposed latent-space aggregation strategy is model-agnostic and transfers across different video diffusion backbones.

7 Conclusion

In this work, we investigate the mechanisms underlying reasoning in diffusion-based video generation models. Contrary to the previously hypothesized Chain-of-Frames (CoF) mechanism, we show that reasoning primarily unfolds along diffusion steps, which we term Chain-of-Steps (CoS), through qualitative analysis and targeted perturbation experiments. Our study further reveals several emergent reasoning behaviors, including working memory retention, self-correction, and perception before action. Analysis of Diffusion Transformer layers shows that the middle layers are important for reasoning. To further validate these insights, we propose a simple training-free reasoning path ensemble method that achieves performance improvements over a strong baseline.

References

1. Alayrac, J.B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., et al.: Flamingo: a visual language model for few-shot learning. Advances in neural information processing systems **35**, 23716–23736 (2022) 4
2. Bai, J., Bai, S., Yang, S., Wang, S., Tan, S., Wang, P., Lin, J., Zhou, C., Zhou, J.: Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. arXiv preprint arXiv:2308.12966 (2023) 4
3. Behrens, T.E., Muller, T.H., Whittington, J.C., Mark, S., Baram, A.B., Stachenfeld, K.L., Kurth-Nelson, Z.: What is a cognitive map? organizing knowledge for flexible behavior. Neuron **100**(2), 490–509 (2018) 5
4. Chen, J., Xu, Z., Pan, X., Hu, Y., Qin, C., Goldstein, T., Huang, L., Zhou, T., Xie, S., Savarese, S., et al.: Blip3-o: A family of fully open unified multimodal models-architecture, training and dataset. arXiv preprint arXiv:2505.09568 (2025) 4
5. Chen, L.L., Ma, H., Fan, Z., Huang, Z., Sinha, A., Dai, X., Wang, J., He, Z., Yang, J., Li, C., et al.: Unit: Unified multimodal chain-of-thought test-time scaling. arXiv preprint arXiv:2602.12279 (2026) 4
6. Chen, Z., Wu, J., Wang, W., Su, W., Chen, G., Xing, S., Zhong, M., Zhang, Q., Zhu, X., Lu, L., et al.: Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 24185–24198 (2024) 4
7. Chern, E., Hu, Z., Chern, S., Kou, S., Su, J., Ma, Y., Deng, Z., Liu, P.: Thinking with generated images. arXiv preprint arXiv:2505.22525 (2025) 4
8. Deng, C., Zhu, D., Li, K., Gou, C., Li, F., Wang, Z., Zhong, S., Yu, W., Nie, X., Song, Z., et al.: Emerging properties in unified multimodal pretraining. arXiv preprint arXiv:2505.14683 (2025) 4
9. Denton, E., Fergus, R.: Stochastic video generation with a learned prior. In: International conference on machine learning. pp. 1174–1183. PMLR (2018) 4
10. Duan, C., Fang, R., Wang, Y., Wang, K., Huang, L., Zeng, X., Li, H., Liu, X.: Got-r1: Unleashing reasoning capability of mllm for visual generation with reinforcement learning. arXiv preprint arXiv:2505.17022 (2025) 4
11. Fan, W., Si, C., Song, J., Yang, Z., He, Y., Zhuo, L., Huang, Z., Dong, Z., He, J., Pan, D., et al.: Vchitect-2.0: Parallel transformer for scaling up video diffusion models. arXiv preprint arXiv:2501.08453 (2025) 4
12. Fan, X., Qiu, Z., Wu, Z., Wang, F., Lin, Z., Ren, T., Lin, D., Gong, R., Yang, L.: Phased dmd: Few-step distribution matching distillation via score matching within subintervals. arXiv preprint arXiv:2510.27684 (2025) 27
13. Fang, R., Duan, C., Wang, K., Huang, L., Li, H., Yan, S., Tian, H., Zeng, X., Zhao, R., Dai, J., et al.: Got: Unleashing reasoning capability of multimodal large language model for visual generation and editing. arXiv preprint arXiv:2503.10639 (2025) 4
14. Ge, Y., Zhao, S., Zhu, J., Ge, Y., Yi, K., Song, L., Li, C., Ding, X., Shan, Y.: Seed-x: Multimodal models with unified multi-granularity comprehension and generation. arXiv preprint arXiv:2404.14396 (2024) 4
15. Google DeepMind: Veo 3.1: Ingredients to Video. Technical Report Veo 3.1, Google DeepMind (January 2026), https://blog.google/innovation-and-ai/technology/ai/veo-3-1-ingredients-to-video/, released January 13, 2026 4, 15

16. Guo, Z., Chen, X., Zhang, R., An, R., Qi, Y., Jiang, D., Li, X., Zhang, M., Li, H., Heng, P.A.: Are video models ready as zero-shot reasoners? an empirical study with the mme-cof benchmark. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9175–9184 (2026) 15
17. Guo, Z., Zhang, R., Tong, C., Zhao, Z., Gao, P., Li, H., Heng, P.A.: Can we generate images with cot? let's verify and reinforce image generation step by step. arXiv preprint arXiv:2501.13926 (2025) 4
18. HaCohen, Y., Brazowski, B., Chiprut, N., Bitterman, Y., Kvochko, A., Berkowitz, A., Shalem, D., Lifschitz, D., Moshe, D., Porat, E., Richardson, E., Shiran, G., Chachy, I., Chetboun, J., Finkelson, M., Kupchick, M., Zabari, N., Guetta, N., Kotler, N., Bibi, O., Gordon, O., Panet, P., Benita, R., Armon, S., Kulikov, V., Inger, Y., Shiftan, Y., Melumian, Z., Farbman, Z.: Ltx-2: Efficient joint audio-visual foundation model (2026), https://arxiv.org/abs/2601.03233, submitted 6 Jan 2026 5, 15
19. Hao, S., Sukhbaatar, S., Su, D., Li, X., Hu, Z., Weston, J., Tian, Y.: Training large language models to reason in a continuous latent space. arXiv preprint arXiv:2412.06769 (2024) 4
20. He, X., Fan, Z., Li, H., Zhuo, F., Xu, H., Cheng, S., Weng, D., Liu, H., Ye, C., Wu, B.: Ruler-bench: Probing rule-based reasoning abilities of next-level video generation models for vision foundation intelligence. arXiv preprint arXiv:2512.02622 (2025) 4
21. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. In: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., Lin, H. (eds.) Advances in Neural Information Processing Systems. vol. 33, pp. 6840–6851. Curran Associates, Inc. (2020), https://proceedings.neurips.cc/paper_files/paper/2020/file/4c5bcfec8584af0d967f1ab10179ca4b-Paper.pdf 4
22. Huang, J., Gu, S., Hou, L., Wu, Y., Wang, X., Yu, H., Han, J.: Large language models can self-improve. In: Proceedings of the 2023 conference on empirical methods in natural language processing. pp. 1051–1068 (2023) 4, 9
23. Huang, Z., He, Y., Yu, J., Zhang, F., Si, C., Jiang, Y., Zhang, Y., Wu, T., Jin, Q., Chanpaisit, N., Wang, Y., Chen, X., Wang, L., Lin, D., Qiao, Y., Liu, Z.: VBench: Comprehensive benchmark suite for video generative models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2024) 6
24. Huang, Z., Yu, N., Chen, G., Qiu, H., Debevec, P., Liu, Z.: Vchain: Chain-of-visual-thought for reasoning in video generation. arXiv preprint arXiv:2510.05094 (2025) 4
25. Huang, Z., Zhang, F., Xu, X., He, Y., Yu, J., Dong, Z., Ma, Q., Chanpaisit, N., Si, C., Jiang, Y., et al.: Vbench++: Comprehensive and versatile benchmark suite for video generative models. IEEE Transactions on Pattern Analysis and Machine Intelligence (2025) 6
26. Jiang, D., Guo, Z., Zhang, R., Zong, Z., Li, H., Zhuo, L., Yan, S., Heng, P.A., Li, H.: T2i-r1: Reinforcing image generation with collaborative semantic-level and token-level cot. arXiv preprint arXiv:2505.00703 (2025) 4
27. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114 (2013) 4
28. Koh, J.Y., Fried, D., Salakhutdinov, R.: Generating images with multimodal language models. NeurIPS (2023) 4
29. Kong, W., Tian, Q., Zhang, Z., Min, R., Dai, Z., Zhou, J., Xiong, J., Li, X., Wu, B., Zhang, J., et al.: HunyuanVideo: A systematic framework for large video generative models. arXiv preprint arXiv:2412.03603 (2024) 4, 15

30. Kornblith, S., Norouzi, M., Lee, H., Hinton, G.: Similarity of neural network representations revisited. In: International conference on machine learning. pp. 3519–3529. PMlR (2019) 9
31. Kuaishou Technology: Kling ai launches video 2.6 model with "simultaneous audio-visual generation" capability, redefining ai video creation workflow. Press Release (December 2025), model released December 3, 2025; Press release published December 5, 2025 4, 15
32. Li, C., Wu, W., Zhang, H., Xia, Y., Mao, S., Dong, L., Vulić, I., Wei, F.: Imagine while reasoning in space: Multimodal visualization-of-thought. arXiv preprint arXiv:2501.07542 (2025) 4
33. Li, J., Li, D., Savarese, S., Hoi, S.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In: International conference on machine learning. pp. 19730–19742. PMLR (2023) 4
34. Li, T., Lu, Q., Zhao, L., Li, H., Zhu, X., Qiao, Y., Zhang, J., Shao, W.: Unifork: Exploring modality alignment for unified multimodal understanding and generation. arXiv preprint arXiv:2506.17202 (2025) 4
35. Liao, C., Liu, L., Wang, X., Luo, Z., Zhang, X., Zhao, W., Wu, J., Li, L., Tian, Z., Huang, W.: Mogao: An omni foundation model for interleaved multi-modal generation. arXiv preprint arXiv:2505.05472 (2025) 4
36. Lin, H., Pan, X., Huang, Z., Hou, J., Wang, J., Chen, W., He, Z., Juefei-Xu, F., Sun, J., Fan, Z., et al.: Exploring mllm-diffusion information transfer with metacanvas. arXiv preprint arXiv:2512.11464 (2025) 4
37. Lipman, Y., Chen, R.T., Ben-Hamu, H., Nickel, M., Le, M.: Flow matching for generative modeling. In: International Conference on Learning Representations (ICLR) (2023) 5
38. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning (2023) 4
39. Liu, M., Ma, S., Meng, S., Zhao, X., Zhang, Z., Zhang, S., Zhong, Z., Chen, P., Cao, H., Sun, X., et al.: Rise-video: Can video generators decode implicit world rules? arXiv preprint arXiv:2602.05986 (2026) 15
40. Luo, Y., Zhao, X., Lin, B., Zhu, L., Tang, L., Liu, Y., Chen, Y.C., Qian, S., Wang, X., You, Y.: V-reasonbench: Toward unified reasoning benchmark suite for video generation models. arXiv preprint arXiv:2511.16668 (2025) 4
41. Luo, Y., Zhao, X., Lin, B., Zhu, L., Tang, L., Liu, Y., Chen, Y.C., Qian, S., Wang, X., You, Y.: V-reasonbench: Toward unified reasoning benchmark suite for video generation models. arXiv preprint arXiv:2511.16668 (2025) 15
42. Madaan, A., Tandon, N., Gupta, P., Hallinan, S., Gao, L., Wiegreffe, S., Alon, U., Dziri, N., Prabhumoye, S., Yang, Y., Welleck, S., Majumder, B.P., Gupta, S., Yazdanbakhsh, A., Clark, P.: Self-refine: Iterative refinement with self-feedback (2023) 4, 9
43. Mattar, M.G., Lengyel, M.: Planning in the brain. Neuron **110**(6), 914–934 (2022) 5
44. OpenAI: Sora: Openai's text-to-video model (September 2025), https://openai.com/index/sora-is-here, publicly released September 2025 4, 15
45. Pan, X., Shukla, S.N., Singh, A., Zhao, Z., Mishra, S.K., Wang, J., Xu, Z., Chen, J., Li, K., Juefei-Xu, F., et al.: Transfer between modalities with metaqueries. arXiv preprint arXiv:2504.06256 (2025) 4
46. Peebles, W., Xie, S.: Scalable diffusion models with transformers. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4195–4205 (2023) 4
47. Pfeiffer, B.E., Foster, D.J.: Hippocampal place-cell sequences depict future paths to remembered goals. Nature **497**(7447), 74–79 (2013) 4

48. Qin, L., Gong, J., Sun, Y., Li, T., Yang, M., Yang, X., Qu, C., Tan, Z., Li, H.: Uni-cot: Towards unified chain-of-thought reasoning across text and vision. arXiv preprint arXiv:2508.05606 (2025) 4
49. Qu, L., Zhang, H., Liu, Y., Wang, X., Jiang, Y., Gao, Y., Ye, H., Du, D.K., Yuan, Z., Wu, X.: Tokenflow: Unified image tokenizer for multimodal understanding and generation. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 2545–2555 (2025) 4
50. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022) 4
51. Runway Research: Introducing runway gen-4: Next-generation ai models for media generation and world consistency (March 2025), https://runwayml.com/research/introducing-runway-gen-4, accessed January 27, 2026 4, 15
52. Shi, C., Yu, Y., Yang, S.: Vision function layer in multimodal llms. arXiv preprint arXiv:2509.24791 (2025) 11
53. Shi, W., Han, X., Zhou, C., Liang, W., Lin, X.V., Zettlemoyer, L., Yu, L.: Lmfusion: Adapting pretrained language models for multimodal generation. arXiv preprint arXiv:2412.15188 (2024) 4
54. Shi, W., Yu, A., Fang, R., Ren, H., Wang, K., Zhou, A., Tian, C., Fu, X., Hu, Y., Lu, Z., et al.: Mathcanvas: Intrinsic visual chain-of-thought for multimodal mathematical reasoning. arXiv preprint arXiv:2510.14958 (2025) 4
55. Su, J., Lu, Y., Pan, S., Wen, B., Liu, Y.: Roformer: Enhanced transformer with rotary position embedding (2021) 24
56. Tan, Z., Yang, H., Qin, L., Gong, J., Yang, M., Li, H.: Omni-video: Democratizing unified video understanding and generation. arXiv preprint arXiv:2507.06119 (2025) 4
57. Team, C.: Chameleon: Mixed-modal early-fusion foundation models. arXiv preprint arXiv:2405.09818 (2024). https://doi.org/10.48550/arXiv.2405.09818, https://github.com/facebookresearch/chameleon 4
58. Tong, J., Mou, Y., Li, H., Li, M., Yang, Y., Zhang, M., Chen, Q., Liang, T., Hu, X., Zheng, Y., et al.: Thinking with video: Video generation as a promising multimodal reasoning paradigm. arXiv preprint arXiv:2511.04570 (2025) 4
59. Tong, S., Fan, D., Li, J., Xiong, Y., Chen, X., Sinha, K., Rabbat, M., LeCun, Y., Xie, S., Liu, Z.: Metamorph: Multimodal understanding and generation via instruction tuning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 17001–17012 (2025) 4
60. Wan, T., Wang, A., Ai, B., Wen, B., Mao, C., Xie, C.W., Chen, D., Yu, F., Zhao, H., Yang, J., et al.: Wan: Open and advanced large-scale video generative models. arXiv preprint arXiv:2503.20314 (2025) 4, 5, 15
61. Wang, M., Wang, R., Lin, J., Ji, R., Wiedemer, T., Gao, Q., Luo, D., Qian, Y., Huang, L., Hong, Z., Ge, J., Ma, Q., He, H., Zhou, Y., Guo, L., Mei, L., Li, J., Xing, H., Zhao, T., Yu, F., Xiao, W., Jiao, Y., Hou, J., Zhang, D., Xu, P., Zhong, B., Zhao, Z., Fang, G., Kitaoka, J., Xu, Y., Xu, H., Blacutt, K., Nguyen, T., Song, S., Sun, H., Wen, S., He, L., Wang, R., Wang, Y., Yang, M., Ma, Z., Millière, R., Shi, F., Vasconcelos, N., Khashabi, D., Yuille, A., Du, Y., Liu, Z., Lin, D., Liu, Z., Kumar, V., Li, Y., Yang, L., Cai, Z., Deng, H.: A very big video reasoning suite. arXiv preprint arXiv:2602.20159 (2026), https://arxiv.org/abs/2602.20159 2, 4, 5, 6, 15, 22
62. Wang, Y., Chen, X., Ma, X., Zhou, S., Huang, Z., Wang, Y., Yang, C., He, Y., Yu, J., Yang, P., et al.: Lavie: High-quality video generation with cascaded latent diffu-

sion models. International Journal of Computer Vision **133**(5), 3059–3078 (2025) 4
63. Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., et al.: Emergent abilities of large language models. arXiv preprint arXiv:2206.07682 (2022) 4
64. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q.V., Zhou, D., et al.: Chain-of-thought prompting elicits reasoning in large language models. Advances in neural information processing systems **35**, 24824–24837 (2022) 4, 9
65. Wiedemer, T., Li, Y., Vicol, P., Gu, S.S., Matarese, N., Swersky, K., Kim, B., Jaini, P., Geirhos, R.: Video models are zero-shot learners and reasoners. arXiv preprint arXiv:2509.20328 (2025) 2, 4, 5
66. Wortsman, M., Ilharco, G., Gadre, S.Y., Roelofs, R., Gontijo-Lopes, R., Morcos, A.S., Namkoong, H., Farhadi, A., Carmon, Y., Kornblith, S., et al.: Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In: International conference on machine learning. pp. 23965–23998. PMLR (2022) 14
67. Wu, C., Chen, X., Wu, Z., Ma, Y., Liu, X., Pan, Z., Liu, W., Xie, Z., Yu, X., Ruan, C., et al.: Janus: Decoupling visual encoding for unified multimodal understanding and generation. arXiv preprint arXiv:2410.13848 (2024) 4
68. Wu, C., Zheng, P., Yan, R., Xiao, S., Luo, X., Wang, Y., Li, W., Jiang, X., Liu, Y., Zhou, J., et al.: Omnigen2: Exploration to advanced multimodal generation. arXiv preprint arXiv:2506.18871 (2025) 4
69. Wu, J.Z., Ren, X., Shen, T., Cao, T., He, K., Lu, Y., Gao, R., Xie, E., Lan, S., Alvarez, J.M., Gao, J., Fidler, S., Wang, Z., Ling, H.: Chronoedit: Towards temporal reasoning for image editing and world simulation. arXiv preprint arXiv:2510.04290 (2025) 2, 24
70. Xiao, Y., Song, L., Chen, Y., Luo, Y., Chen, Y., Gan, Y., Huang, W., Li, X., Qi, X., Shan, Y.: Mindomni: Unleashing reasoning generation in vision language models with rgpo. arXiv preprint arXiv:2505.13031 (2025) 4
71. Xu, Y., Li, C., Zhou, H., Wan, X., Zhang, C., Korhonen, A., Vulić, I.: Visual planning: Let's think only with images (2025), https://arxiv.org/abs/2505.11409 4
72. Yang, S., Wu, J., Chen, X., Xiao, Y., Yang, X., Wong, D.F., Wang, D.: Understanding aha moments: from external observations to internal mechanisms. arXiv preprint arXiv:2504.02956 (2025) 4, 9, 11
73. Yang, Z., Teng, J., Zheng, W., Ding, M., Huang, S., Xu, J., Yang, Y., Hong, W., Zhang, X., Feng, G., et al.: CogVideoX: Text-to-video diffusion models with an expert transformer. arXiv preprint arXiv:2408.06072 (2024) 15
74. Yao, S., Yu, D., Zhao, J., Shafran, I., Griffiths, T., Cao, Y., Narasimhan, K.: Tree of thoughts: Deliberate problem solving with large language models. In: Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., Levine, S. (eds.) Advances in Neural Information Processing Systems. vol. 36, pp. 11809–11822. Curran Associates, Inc. (2023), https://proceedings.neurips.cc/paper_files/paper/2023/file/271db9922b8d1f4dd7aaef84ed5ac703-Paper-Conference.pdf 7
75. Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K., Cao, Y.: ReAct: Synergizing reasoning and acting in language models. In: International Conference on Learning Representations (ICLR) (2023) 4, 9
76. Yue, J., Huang, Z., Chen, Z., Wang, X., Wan, P., Liu, Z.: Simulating the visual world with artificial intelligence: A roadmap. arXiv preprint arXiv:2511.08585 (2025) 4

77. Zawalski, M., Chen, W., Pertsch, K., Mees, O., Finn, C., Levine, S.: Robotic control via embodied chain-of-thought reasoning. arXiv preprint arXiv:2407.08693 (2024) 4
78. Zeng, S., Chang, X., Xie, M., Liu, X., Bai, Y., Pan, Z., Xu, M., Wei, X.: Futuresightdrive: Thinking visually with spatio-temporal cot for autonomous driving. arXiv preprint arXiv:2505.17685 (2025) 4
79. Zhao, Q., Lu, Y., Kim, M.J., Fu, Z., Zhang, Z., Wu, Y., Li, Z., Ma, Q., Han, S., Finn, C., et al.: Cot-vla: Visual chain-of-thought reasoning for vision-language-action models. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 1702–1713 (2025) 4
80. Zhao, S., Zhang, Y., Cun, X., Yang, S., Niu, M., Li, X., Hu, W., Shan, Y.: Cv-vae: A compatible video vae for latent generative video models. Advances in Neural Information Processing Systems **37**, 12847–12871 (2024) 4
81. Zheng, D., Huang, Z., Liu, H., Zou, K., He, Y., Zhang, F., Zhang, Y., He, J., Zheng, W.S., Qiao, Y., Liu, Z.: VBench-2.0: Advancing video generation benchmark suite for intrinsic faithfulness. arXiv preprint arXiv:2503.21755 (2025) 4
82. Zheng, K., He, X., Wang, X.E.: Minigpt-5: Interleaved vision-and-language generation via generative vokens (2023) 4
83. Zhuang, X., Xie, Y., Deng, Y., Liang, L., Ru, J., Yin, Y., Zou, Y.: Vargpt: Unified understanding and generation in a visual autoregressive multimodal large language model. arXiv preprint arXiv:2501.12327 (2025) 4
84. Zou, K., Huang, Z., Dong, Y., Tian, S., Zheng, D., Liu, H., He, J., Liu, B., Qiao, Y., Liu, Z.: Uni-mmmu: A massive multi-discipline multimodal unified benchmark. arXiv preprint arXiv:2510.13759 (2025) 4