OpenCoF: Learning to Reason Through Video Generation

Xinyan Chen[1,2,*] **Ziyu Guo**[3,*] **Renrui Zhang**[1,†] **Dongzhi Jiang**[1,2] **Hongsheng Li**[2]

[1]ByteDance Seed [2]CUHK MMLab [3]CUHK IMIXR

[*]Equal contribution [†]Corresponding author (renruizhang@bytedance.com)

Abstract

Reasoning has become a core capability for large models, especially when reliable decisions require understanding logical consequences. Recent video generation models offer a reasoning path distinct from previous Chain-of-Thought (CoT): reasoning can unfold through temporally connected frames, known as Chain-of-Frame (CoF) reasoning. However, existing video generators are primarily trained on general video corpora, still lacking diverse supervision and dedicated designs for CoF reasoning. To address this gap, we introduce OpenCoF, a framework comprising the OpenCoF-17K dataset, a reasoning video dataset spanning 11 task families, and Wan-CoF, a fine-tuned video model for studying whether diverse temporal supervision improves CoF behavior. Across four video reasoning benchmarks, Wan-CoF achieves considerable gains over the Wan2.2-I2V-A14B baseline. Building on this, we empirically explore more advanced designs for CoF capabilities, i.e., equipping the model with visual and textual reasoning tokens. This mechanism respectively captures low-level visual cues and high-level semantic priors for spatial and temporal reasoning. Through performance comparisons and attention analysis, we examine how these tokens contribute across model depth, denoising steps, space, and time. Our results suggest that stronger video reasoning requires both broad temporal supervision and explicit mechanisms for organizing intermediate reasoning state. We release the dataset, model, and code. More details are available on our project page: https://opencof.github.io/.

1 Introduction

Enhancing reasoning ability [24, 51, 16, 50] has emerged as a central objective in the development of large language (LLM) and multimodal (LMM) models. Prior works demonstrate that robust reasoning capabilities are indispensable for reliable decision-making in vision-language contexts. However, mainstream visual Chain-of-Thought (CoT) [45, 17, 27] pipelines remain largely anchored to static visual observations. These approaches typically rely on extracting localized evidence [4, 35], invoking external tools [15, 52, 13, 11], or employing auxiliary image-generation steps [21, 48], which limits their ability to intrinsically model dynamic transitions and multi-step visual consequences.

The rapid advancement of video generation models offers a promising alternative: reasoning via video [46, 31, 10, 23]. Rather than relying solely on textual steps or static visual content, models can reason through the temporal evolution of frames, a process formalized as Chain-of-Frame (CoF) reasoning. Recent studies suggest that video models already exhibit nontrivial world knowledge and early signs of physical and causal understanding, indicating significant potential for CoF [46]. Nevertheless, this potential remains far from fully realized. In reasoning-intensive tasks, current video models still struggle with long-term temporal coherence, physical and spatial consistency, and logical continuity [10]. Ultimately, high visual realism does not guarantee reliable reasoning.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

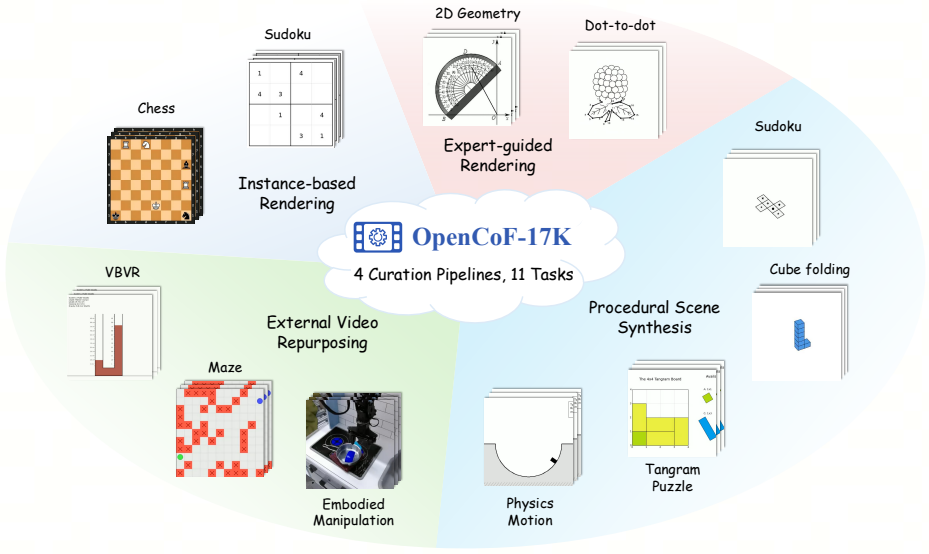


Figure 1: **Overview of OpenCoF.** We construct the OpenCoF-17K dataset, comprising 17,312 videos across 11 tasks, via four complementary pipelines, to provide diverse temporal supervision for Chain-of-Frame (CoF) reasoning.

Starting from MME-CoF [10], recent works have predominantly focused on establishing benchmarks to delineate and evaluate the reasoning scope of video generation models, including RULER-Bench [12], Gen-ViRe [23], VIPER [22], and others [25, 40]. While these efforts are instrumental in defining the problem space, research dedicated to actively enhancing these reasoning capabilities remains noticeably sparse. A few preliminary studies have explored tailored training and test-time scaling strategies, though they often concentrate on specific task domains, such as maze solving in VR-Bench [49] and Thinking in Frames [20], or custom cognitive-task suites like VBVR [44]. Despite these initial advancements, current approaches exhibit two limitations: (1) Their training and evaluation protocols are localized to specific domains or intrinsically coupled to their own curated suites, leaving true generalizability to independent, externally curated reasoning benchmarks underexplored; and (2) their primary emphasis has been on data scaling and inference strategies, leaving internal video model architectures explicitly tailored for reasoning largely uninvestigated. Consequently, there remains a fundamental gap in understanding how to structurally enhance video reasoning capabilities.

To bridge this gap, we introduce OpenCoF, a framework dedicated to advancing video reasoning. As the foundation of this project, we construct the OpenCoF-17K dataset, a diverse collection comprising 17,312 videos across 11 task families. Its curation pipeline integrates four distinct approaches: instance-based rendering, expert-guided rendering, procedural scene synthesis, and the repurposing of existing videos, as illustrated in Figure 1. Leveraging this dataset, we investigate the enhancement of video reasoning capabilities within a robust open-source video generation model, Wan2.2-I2V-A14B [43]. First, we fine-tune it on OpenCoF-17K without incorporating any reasoning-specific techniques, yielding Wan-CoF. This data-centric phase aims to verify whether diverse temporal supervision alone can effectively elicit CoF behavior. The subsequent evaluations across four external video-reasoning benchmarks demonstrate substantial gains over the base model.

Beyond this data-centric stage, we further explore how dedicated reasoning techniques can help a video generator organize the intermediate reasoning state required by CoF. In current DiT-based video generators, such reasoning remains implicit in generation-oriented visual latents and text conditions. We therefore explore two complementary reasoning-token designs that operate at different levels of the model: Visual Reasoning Tokens (*vt*), inserted into the visual latent sequence to capture low-level visual cues, and Textual Reasoning Tokens (*tt*), appended to the text-conditioning sequence to provide high-level semantic priors. These two variants further improve Wan-CoF on external benchmarks while exhibiting different strengths across task dimensions, as shown in Figure 2. Through empirical

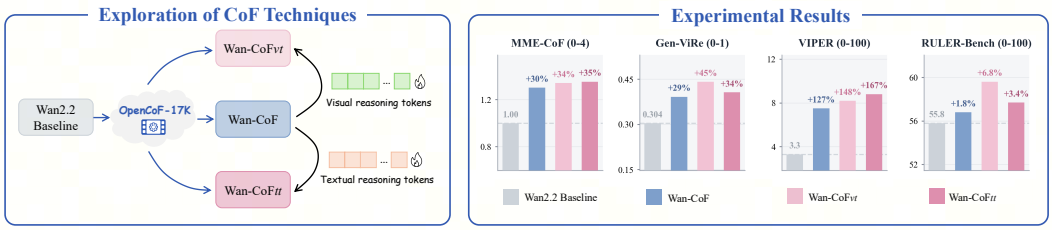


Figure 2: **Exploration of CoF techniques.** Fine-tuning Wan2.2-I2V-A14B on OPENCOF-17K yields WAN-COF, which achieves substantial gains over the Wan2.2 baseline across four external video-reasoning benchmarks (MME-CoF, Gen-ViRe, VIPER, RULER-Bench). We further explore Visual Reasoning Tokens (*vt*) and Textual Reasoning Tokens (*tt*) as complementary reasoning-token designs, yielding WAN-COF$_{vt}$ and WAN-COF$_{tt}$, which improve further while exhibiting different strengths across benchmarks.

performance evaluations and attention analysis, we examine how these token types facilitate reasoning across model depth, denoising steps, spatial dimensions, and time.

Our contributions are summarized as follows:

- We introduce the OPENCOF-17K dataset, comprising 17K videos across 11 task families with a scalable four-source curation pipeline.
- We develop WAN-COF by fine-tuning Wan2.2-I2V-A14B on OPENCOF-17K, demonstrating substantial gains over the base model across four external video-reasoning benchmarks.
- We propose complementary reasoning-token mechanisms (*vt* and *tt*) and study their role in CoF reasoning through performance and attention analyses.

2 OPENCOF-17K

Current video generators lack dedicated temporal supervision for reasoning tasks, making diverse reasoning-centric data essential for robust CoF capabilities. To meet this need, we construct OPENCOF-17K, a video reasoning dataset spanning 11 task families. We first summarize its scale, format, and statistics in Section 2.1, then detail the four-part curation pipeline in Section 2.2.

2.1 Dataset Statistics

The OPENCOF-17K dataset comprises 17,312 samples across 11 distinct task families. Formulated for conditional generation, each data instance pairs an initial conditioning image and a textual prompt with a target reasoning video. To ensure optimization consistency across diverse data sources, all videos are standardized to a resolution of 480p, a frame rate of 15 fps, and a temporal length of 81 frames. Table 1 reports the per-task statistics of OPENCOF-17K. In total the dataset comprises 17,312 videos across 11 task families, with an average prompt length of 53.2 words. The VBVR family aggregates 30 subtasks: 29 of them contribute 250 demonstration videos each and the remaining one contributes 500, summing to 7,750 videos. Figures 8 and 9 show representative cases for each of the 11 task families in the OPENCOF-17K dataset. For every case we display three uniformly sampled frames from the demonstration video together with the original text prompt. We next describe how these data are collected and rendered into a unified video-reasoning format.

2.2 Dataset Curation

To obtain the coverage summarized above, our curation protocol comprises four pipelines: (1) Instance-based rendering, (2) Expert-guided rendering, (3) Procedural scene synthesis, and (4) External video repurposing. As illustrated in Figure 4, the first three pipelines focus on generative construction. To complement these efforts, the fourth pipeline strategically repurposes high-quality demonstrations from existing external datasets. Collectively, these approaches integrate multifaceted data into a unified video-reasoning format, establishing a scalable foundation for future expansion.

Figure 3: Composition of the OpenCoF-17K dataset.

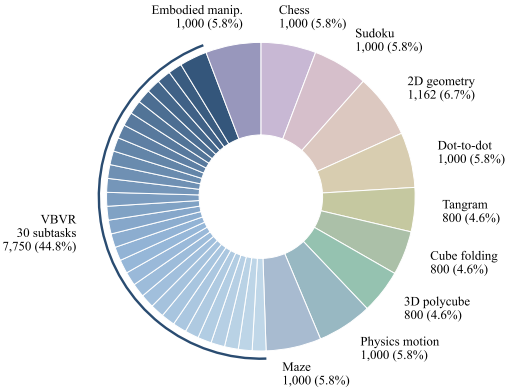


Table 1: Per-task statistics of OpenCoF-17K. `Avg prompt` denotes the mean prompt length in words.

Task	Num	Avg prompt
Chess	1,000	42.0
Sudoku	1,000	52.0
2D geometry	1,162	23.4
Dot-to-dot	1,000	31.0
Tangram puzzle	800	13.5
Cube folding	800	17.7
3D polycube rotation	800	25.3
Physics motion	1,000	45.9
Maze	1,000	179.4
Embodied manipulation	1,000	10.2
VBVR (30 subtasks)	$250 \times 29 + 500$	62.9
Total	**17,312**	**53.2**

In contrast to prior efforts like VBVR and VR-Bench that predominantly rely on procedural scene synthesis, our dataset curation pipeline combines multiple complementary construction strategies.

2.2.1 Instance-based Rendering

Within this pipeline, we synthesize video data from pre-existing structured assets, such as puzzle states or board configurations. Individual instances are sampled from these assets and sequentially translated into video formats via deterministic code rendering.

Chess. This task aims to bolster the reasoning proficiency of video models within highly structured, rule-based environments. We sample instances from the Mate in One (Chess)[1] dataset and procedurally render the corresponding video sequences. The models are tasked with identifying and visually executing the singular legal move that results in a checkmate.

Sudoku. To inject algorithmic and computational reasoning supervision, we leverage the training split of the `Sudoku-Easy`[2] dataset, augmenting it with procedurally generated puzzles of equivalent difficulty. Here, video models are required to resolve 4×4 Sudoku puzzles through a rigorous, step-by-step deductive process.

2.2.2 Expert-guided Rendering

Under this protocol, we leverage human expertise, LLM-generated concepts, or outputs from text-to-image (T2I) models to establish structures and annotations. Guided by these representations, we extract or render intermediate structures and convert them into sequential videos with code.

2D geometry. Targeting geometric reasoning, we sample 2D geometry images from Geo170K [7] that require auxiliary-line construction. Expert annotators define ground-truth coordinates for these segments, after which we render videos showing the step-by-step drawing process.

Dot-to-dot. The dot-to-dot task can strengthen the model's understanding of abstract graphics and its ability to follow strict sequential rules. We predefine various categories and prompt LLMs [6] to generate specific concepts. Subsequently, we synthesize minimalist line-drawing images using T2I models [36], which are then converted into sequential dot-to-dot animation videos via classical computer vision routines.

2.2.3 Procedural Scene Synthesis

This pipeline synthesizes video data entirely from scratch. We utilize random seeds to instantiate foundational scene parameters, while physics engines or graphics pipelines subsequently render the unfolding dynamics into continuous video streams.

[1] https://www.kaggle.com/datasets/ancientaxe/mate-in-one-chess
[2] https://huggingface.co/datasets/LangAGI-Lab/Sudoku-Easy

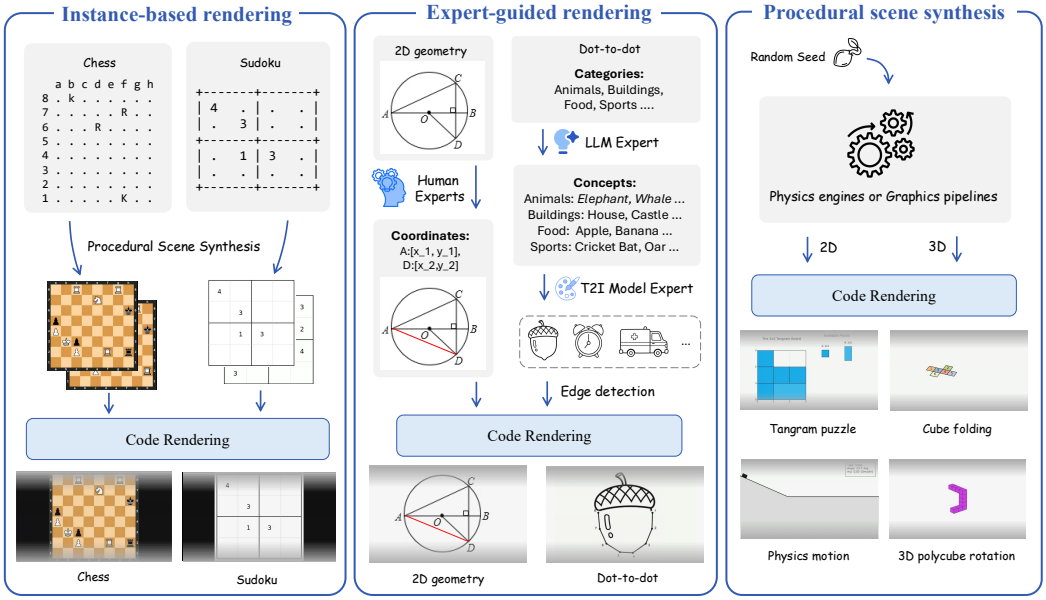


Figure 4: **Overview of the first three curation pipelines in OpenCoF-17K.** Instance-based rendering converts structured assets into videos through code rendering. Expert-guided rendering uses human, LLM, or T2I-model guidance to define structures before rendering. Procedural scene synthesis uses physics engines or graphics pipelines to instantiate scenes.

Tangram puzzle. Designed to cultivate spatial awareness and shape manipulation capabilities, this task requires models to systematically rearrange tangram pieces into a target configuration. We procedurally generate puzzle pieces with randomized shapes and initial placements, directly rendering the sequential assembly process into visual formats.

Cube folding. Targeting 2D-to-3D spatial reasoning, we procedurally generate cube-folding sequences. Given a flat 2D cube net, the models must predict the correct folding sequence to form a complete 3D cube. The initial nets are randomized across varying topologies, viewing angles, and surface patterns, with the folding dynamics programmatically animated using Matplotlib 3D.

3D polycube rotation. To inject geometric motion control supervision across diverse perspectives, we formulate a 3D polycube rotation task. Here, polycubes are instantiated with randomized complex geometries and viewpoints, and are subsequently rotated across varying axes and angles.

Physics motion. Physics-based motion reasoning serves as a capability for grounding video models in physical reality. We simulate scenarios featuring balls and blocks traversing flat or curved tracks under varying friction coefficients (e.g., smooth versus rough surfaces). Models are tasked with forecasting the plausible dynamic trajectories of these objects strictly from their initial states.

2.2.4 External Video Repurposing

Complementing the three active-construction pipelines visualized in Figure 4, we strategically leverage high-quality demonstrations from existing video datasets. These resources are rigorously repurposed by standardizing their spatial resolution, frame rate, and temporal length to seamlessly align with our unified training suite.

Maze. We sample maze navigation data from the training split of the VR-Bench [49] dataset. Since some instances in this dataset follow similar distributions, we sample a subset while preserving coverage across subtasks and difficulty levels.

VBVR. Because some VBVR subtasks are either overly simple or target highly similar capabilities, we select 30 representative subtasks from the broader VBVR [44] dataset and randomly sample demonstration videos from each.

Table 2: **Results on MME-CoF [10].** The scores are reported on a 0–4 scale.

Model	Overall	Instruction Alignment	Temporal Consistency	Visual Stability	Content Fidelity	Focus Relevance
Closed-source Models						
Kling-v1 [19]	0.71	0.09	0.21	2.46	0.30	0.48
Seedance-1.0-Pro [8]	1.48	0.39	1.71	2.05	1.22	2.04
Veo-3.0-Fast [9]	1.52	0.73	1.41	1.83	1.25	2.36
Veo-3.0-Preview [9]	1.53	0.70	1.46	1.91	1.24	2.36
Sora-2 [29]	1.80	0.74	1.60	2.36	1.75	2.53
Open-source Models						
Wan2.2-TI2V-5B [43]	0.72	0.08	0.96	1.16	0.48	0.90
HunyuanVideo-I2V [18]	0.98	0.23	1.38	1.52	0.68	1.08
Data-only Fine-tuning						
Wan2.2-I2V-A14B (Baseline) [43]	1.00	0.22	0.96	1.52	0.68	1.62
WAN-COF	1.30	0.31	1.29	1.99	1.00	1.89
Δ over the Baseline	*+0.30*	*+0.10*	*+0.33*	*+0.47*	*+0.32*	*+0.28*
Exploration of Reasoning Techniques						
WAN-COF$_{vt}$	1.34	0.26	1.38	2.33	0.99	1.76
WAN-COF$_{tt}$	1.35	0.34	1.27	2.18	1.17	1.77

Table 3: **Results on Gen-ViRe [23].** Following the Gen-ViRe protocol, the scores are on a 0–1 scale.

Model	Overall	Abstract	Algo. & Logi.	Analogy	Perceptual	Planning	Spatio-Temporal
Closed-source Models							
Kling-v1 [19]	0.198	0.071	0.057	0.117	0.140	0.443	0.359
Seedance-1.0-Pro [8]	0.301	0.154	0.164	0.083	0.171	0.609	0.621
Veo-3.1 [9]	0.486	0.440	0.451	0.367	0.386	0.722	0.550
Sora-2 [29]	0.560	0.604	0.472	0.483	0.496	0.768	0.537
Data-only Fine-tuning							
Wan2.2-I2V-A14B (Baseline) [43]	0.304	0.220	0.359	0.083	0.190	0.519	0.452
WAN-COF	0.391	0.328	0.506	0.133	0.347	0.578	0.451
Δ over the Baseline	*+0.087*	*+0.107*	*+0.147*	*+0.050*	*+0.157*	*+0.059*	*-0.001*
Exploration of Reasoning Techniques							
WAN-COF$_{vt}$	0.441	0.475	0.507	0.233	0.298	0.628	0.507
WAN-COF$_{tt}$	0.406	0.424	0.474	0.167	0.306	0.565	0.499

Embodied manipulation. The data for the embodied manipulation task are sampled from BridgeData V2 [42]. This task requires video models to comprehend spatial relationships and perform controlled physical manipulations.

3 Training of WAN-COF

Before introducing explicit reasoning techniques in the models, we first conduct the most direct study: how far can a strong open video generator go when it is trained only with diverse CoF supervision? To answer this, we use Wan2.2-I2V-A14B [43], a robust general-domain pretrained I2V model, and directly fine-tune it on OPENCOF-17K with LoRA [14]. The resulting model, WAN-COF, marks the data-only stage of our study and serves as a clean reference point for attributing performance gains to the dataset itself. We evaluate WAN-COF on four video reasoning benchmarks, which allows us to separate the effect of OPENCOF-17K from the reasoning-token designs explored later in Section 4. Please refer to Section A.3 for more experimental details.

3.1 Main Results

Comparison with the baseline. As shown in the Data-only Fine-tuning rows of Tables 2 to 5, relative to our baseline, Wan2.2-I2V-A14B, WAN-COF lifts the headline metric on every benchmark: MME-CoF Overall from 1.00 to 1.30 (+0.30), Gen-ViRe average from 0.304 to 0.391 (+0.087), VIPER POC from 3.3 to 7.5 (+4.2), and RULER-Bench overall average from 55.8 to 56.8 (+1.0). The gains are broad-based—every MME-CoF aspect improves, and on Gen-ViRe and VIPER almost every category also gains. The severe non-positive cells sit within evaluation noise or correspond to dimensions where open-source models still score near zero. Moreover, the largest jumps cluster on reasoning-oriented sub-dimensions like Temporal Consistency, Physics, Spatial, Perceptual, and

Table 4: **Results on VIPER [22].** We follow the original VIPER protocol and report POC@1.0.

Model	Overall	Temporal	Structural	Symbolic	Spatial	Physics	Planning
Closed-source Models							
Seedance-1.5 [34]	9.5	5.6	5.6	0.0	5.3	6.5	33.8
Veo-3.1 [9]	20.3	22.2	20.0	13.3	10.7	14.7	41.0
Sora-2 [29]	23.3	10.4	42.5	58.3	4.0	9.4	15.1
Open-source Models							
Hunyuan-1.5 [47]	8.1	8.4	5.0	0.0	2.7	11.1	21.1
Data-only Fine-tuning							
Wan2.2-I2V-A14B (Baseline) [43]	3.3	5.6	6.2	0.0	0.0	2.8	5.3
WAN-COF	7.5	18.8	10.0	0.0	5.3	8.3	5.3
Δ over the Baseline	*+4.2*	*+13.2*	*+3.8*	*0.0*	*+5.3*	*+5.5*	*0.0*
Exploration of Reasoning Techniques							
WAN-COF$_{vt}$	8.2	12.8	11.3	0.0	4.0	8.3	15.8
WAN-COF$_{tt}$	8.8	14.8	15.0	0.0	8.0	2.8	10.5

Table 5: **Results on RULER-Bench [12].** We report the filtered image-to-video setting on a 0–100 scale, since our baseline does not support text-to-video generation. IF: Instruction Following, VC: Visual Consistency, VF: Visual Fidelity, RC: Rule Coherence.

Model	Overall	Vision					Science					Game					Humanity				
		IF	VC	VF	RC	Avg.	IF	VC	VF	RC	Avg.	IF	VC	VF	RC	Avg.	IF	VC	VF	RC	Avg.
Data-only Fine-tuning																					
Wan2.2-I2V-A14B (Baseline) [43]	55.8	0.0	69.1	65.8	28.0	54.3	44.2	73.5	83.6	18.4	54.9	21.6	65.6	83.4	12.0	45.6	61.1	91.7	91.7	40.2	71.2
WAN-COF	56.8	0.0	65.1	67.0	31.7	54.6	45.6	70.8	83.1	18.9	54.6	25.0	74.4	87.0	23.9	52.6	47.2	86.1	94.4	49.4	69.3
Δ over the Baseline	*+1.0*	*0.0*	*-4.0*	*+1.3*	*+3.6*	*+0.3*	*+1.4*	*-2.7*	*-0.5*	*+0.5*	*-0.3*	*+3.4*	*+8.8*	*+3.6*	*+11.9*	*+6.9*	*-13.9*	*-5.6*	*+2.8*	*+9.3*	*-1.9*
Exploration of Reasoning Techniques																					
WAN-COF$_{vt}$	59.6	0.0	72.7	71.0	35.6	59.8	48.5	68.5	81.4	15.5	53.5	30.7	73.4	86.4	22.1	53.2	58.3	88.9	98.2	56.9	75.6
WAN-COF$_{tt}$	58.4	0.0	71.2	68.1	34.2	57.8	45.9	69.8	84.6	18.9	54.8	28.6	76.5	87.2	23.1	53.8	52.8	86.1	94.9	49.4	70.8

Algorithmic & Logical, rather than on visual fidelity alone, suggesting that our training strengthens frame-level reasoning rather than acting as a visual-quality regularizer.

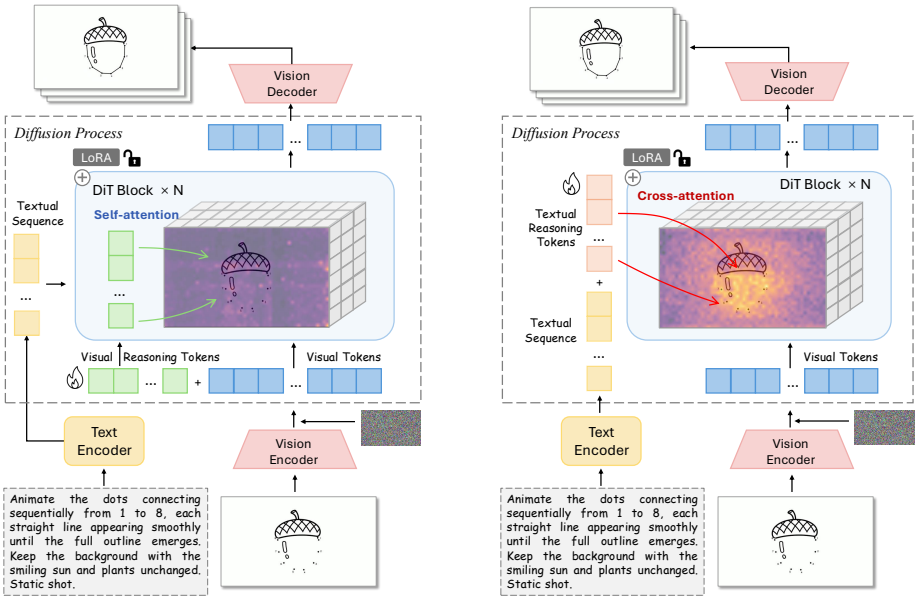
Comparison with other models. In the Data-only Fine-tuning rows of Tables 2 to 5, WAN-COF surpasses open-source competitors and narrows the gap to the closed-source frontier. On MME-CoF, WAN-COF exceeds HunyuanVideo-I2V and Wan2.2-TI2V-5B, approaching Seedance-1.0-Pro. On Gen-ViRe, WAN-COF reaches 0.391 average, outperforming Kling-v1 and Seedance-1.0-Pro and moving closer to Veo-3.1. On VIPER, WAN-COF attains POC of 7.5, on par with open-source Hunyuan-1.5 and closed-source Seedance-1.5. Under a LoRA fine-tuning budget on a single open backbone, this places WAN-COF among the strongest open-source I2V models on CoF reasoning.

3.2 Out-of-distribution Analysis

A key challenge in evaluating CoF reasoning is ensuring that models generalize beyond their training distributions. While recent works [49, 44] evaluate distribution shifts internally within their respective task suites, we explore generalization by transferring our model directly to independent, externally curated benchmarks. Because these external benchmarks encompass distributions distinct from our training set, testing on them inherently measures performance under a distribution shift. Under this setting, WAN-COF demonstrates consistent improvements across all four benchmarks. Notably, the per-category gains appear to align with the types of supervision provided by OPENCOF-17K: training on Physics motion transfers to VIPER Physics (+5.5); structured-grid tasks (e.g., Chess, Sudoku, Maze) correlate with Gen-ViRe Algorithmic & Logical (+0.147) and the RULER-Bench Game split (+6.9); and spatial-shape tasks map to Gen-ViRe Perceptual (+0.157) and VIPER Spatial (+5.3). Together, these transfer signals suggest that OPENCOF-17K helps foster transferable CoF reasoning skills across suite boundaries, rather than overfitting to specific training domains. We further provide a task-level decomposition into task types overlapping with or unseen during OPENCOF-17K fine-tuning in Section A.4.2.

4 Exploration of CoF Techniques

The data-only experiment in Section 3 establishes an encouraging starting point: training on OPENCOF-17K transfers to external reasoning benchmarks even when reasoning techniques are not explicitly added. This result also sharpens the empirical question: if CoF gains can emerge from



(a) Visual Reasoning Tokens **(b) Textual Reasoning Tokens**

Figure 5: **Overview of reasoning-token designs.** (a) Visual Reasoning Tokens are inserted into the visual token sequence and interact with visual tokens through self-attention. (b) Textual Reasoning Tokens are appended to the text-conditioning sequence and guide generation through cross-attention.

supervision alone, can dedicated reasoning techniques give the model a more explicit place to organize the intermediate rules, plans, and abstract goals behind those gains? The underlying generator still lacks a dedicated channel for maintaining reasoning state, and must express such information implicitly through pixel-level latents. Since existing video generation models are not originally designed for stepwise reasoning, they offer no specialized structures to host either the low-level reasoning cues operating inside the visual latent space or the high-level reasoning semantics operating at the prompt level. We therefore explore two complementary designs to target these distinct layers: Visual Reasoning Tokens (*vt*), which reside within the latent space of the DiT to capture low-level reasoning cues, and Textual Reasoning Tokens (*tt*), which augment the text condition to capture high-level reasoning semantics. We subsequently analyze *when, where, and how such tokens help*.

4.1 Two Strategies of Reasoning Tokens

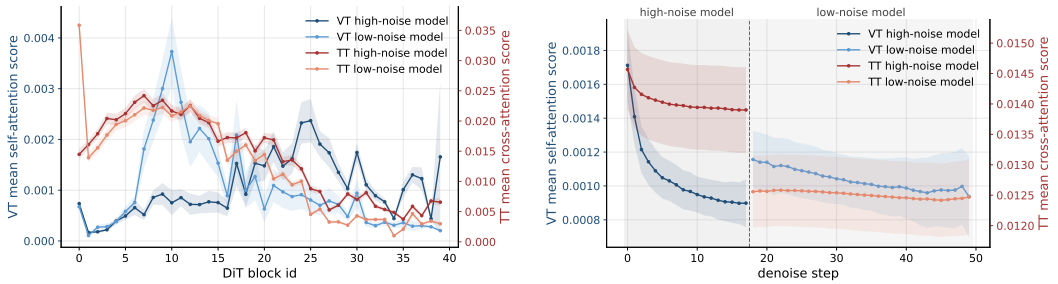
We now introduce our two reasoning-token strategies, as illustrated in Figure 5. The formulation is agnostic to the specific video generator and applies to any DiT-based architecture. For concreteness we instantiate it on Wan2.2-I2V-A14B, the same backbone used for WAN-COF.

Visual reasoning tokens (*vt*). A DiT video generator takes as input a sequence of visual tokens $x \in \mathbb{R}^{T\times H\times W\times d}$ that represent the compressed latent of the video, flattens them into a 1D sequence, and processes them through a stack of self-attention blocks. To host reasoning state directly in this visual space, we introduce N_r learnable Visual Reasoning Tokens, denoted $r^v \in \mathbb{R}^{N_r\times d}$, initialized randomly and prepended to the visual token sequence before the first DiT block:

$$z_0 = [r^v_1, r^v_2, \ldots, r^v_{N_r}, x_1, x_2, \ldots, x_M], \tag{1}$$

where $M = T \times H \times W$ is the total number of visual tokens. Inside every DiT block, self-attention lets each visual token x_i attend to every reasoning token r^v_j, and the reasoning tokens in turn attend to the full visual sequence and to each other. This bidirectional flow enables *vt* to aggregate global context and distribute the resulting features back to the visual tokens. After the final DiT block, the reasoning tokens are discarded and the prediction is read from the visual portion of the output.

Textual reasoning tokens (*tt*). A complementary place to host reasoning state is the text-conditioning side. The DiT consumes a text-conditioning sequence $c \in \mathbb{R}^{L\times d}$ produced by a



(a) Depth-wise attention

(b) Denoising-step attention

Figure 6: **Attention patterns of reasoning tokens.** We average attention over 120 MME-CoF cases during inference. For *vt*, we report visual-token attention to *vt* in self-attention; for *tt*, we report visual-token attention to *tt* in cross-attention. Panel (a) shows depth-wise attention, and panel (b) shows denoising-step attention.

text encoder, where L is the length of that sequence. We introduce N_t learnable Textual Reasoning Tokens, denoted $r^t \in \mathbb{R}^{N_t \times d}$, and prepend them to this sequence before it enters the cross-attention of each DiT block:

$$c_0 = [r_1^t, r_2^t, \ldots, r_{N_t}^t, c_1, c_2, \ldots, c_L]. \tag{2}$$

The cross-attention in each block then lets every visual token x_i attend to both the original text tokens and the new r_j^t, so r^t supplies task-level information that supplements the per-prompt content of c. Unlike *vt*, r^t enters only as additional key/value context and never appears on the query side, so it is not refreshed by the visual sequence and is shared, unchanged, across every spatial patch and every temporal frame. This also means r^t does not need to be removed before the final output readout. *tt* therefore acts as a unidirectional, prompt-independent textual prior, complementary to the spatially grounded, bidirectional reasoning channel that *vt* places inside the visual stream.

4.2 Main Results

We fine-tune Wan2.2-I2V-A14B with LoRA using *vt* and *tt* separately, yielding WAN-COF$_{vt}$ and WAN-COF$_{tt}$. As shown in the Exploration of Reasoning Techniques rows of Tables 2 to 5, both variants improve further over WAN-COF on the headline metric of every benchmark. WAN-COF$_{vt}$ reaches the highest aggregate score on Gen-ViRe and RULER-Bench, while WAN-COF$_{tt}$ leads on MME-CoF and VIPER. At the sub-dimension level, the two variants specialize in opposite directions, in a way consistent with their design: *vt* dominates dimensions that require maintaining a global, persistent plan inside the visual sequence, such as Planning in VIPER, Abstract, Analogy and Planning in Gen-ViRe, and Visual Stability in MME-CoF. In contrast, *tt* dominates dimensions that benefit from a learnable, prompt-independent textual prior, such as Instruction Alignment in MME-CoF and Structural in VIPER.

4.3 Attention Analysis

To understand how the tokens operate, we extract attention during inference on 120 MME-CoF cases and average two signals: visual tokens' self-attention [41] scores assigned to *vt* and cross-attention [41] scores assigned to *tt*. We then inspect these attention patterns along four axes: DiT block depth, denoising step, spatial position within frames, and temporal position across frames.

Depth-wise attention. Panel (a) of Figure 6 shows that both *vt* and *tt* are used non-uniformly across DiT depth, suggesting that reasoning tokens mainly participate in specific computation stages rather than acting as constant global biases. *vt* shows stronger layer selectivity: in the low-noise model it peaks around shallow-to-middle blocks, while in the high-noise model it remains active into later blocks and exhibits stronger mid-to-late spikes, consistent with visual reasoning tokens helping organize latent spatial-temporal structure when denoising is harder. *tt* follows a smoother early-to-middle emphasis and then decays with depth, matching its role as a prompt-level prior supplied through cross-attention.

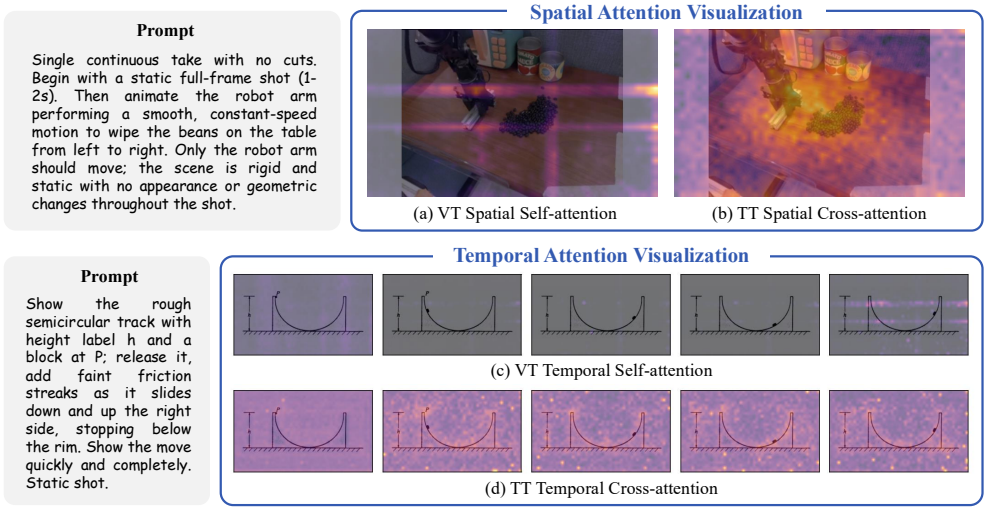


Figure 7: **Spatial and temporal attention visualizations.** Top: representative high-noise spatial attention maps for *vt* self-attention and *tt* cross-attention on an embodied manipulation prompt. Bottom: temporal attention maps for *vt* and *tt* on a physics-reasoning prompt, where frames are uniformly sampled over time and ordered from left to right.

Denoising-step attention. Panel (b) of Figure 6 shows a similar pattern along the denoising trajectory. Both token types receive stronger attention early in denoising, especially in the high-noise model, where global semantics and motion structure are still underdetermined. *vt* attention starts high and drops quickly in the high-noise model, indicating an early role in visual latent planning; in the low-noise model, it becomes flatter and lower, consistent with later refinement. *tt* is smoother across denoising steps, but its high-noise curve stays above the low-noise curve, suggesting that task-level textual priors are most useful before the generation process settles.

Spatial attention. Within a single frame, RT attention reveals different spatial roles for *vt* and *tt*, as visualized in the top row of Figure 7. *vt* attention is relatively sparse and forms pronounced horizontal and vertical bands, with stronger responses near the gripper and the beans. This pattern suggests that *vt* captures localized visual cues while retaining axis-aligned structure from the visual token grid, possibly influenced by Wan's separable 3D RoPE [39]. In contrast, *tt* attention is more diffuse and covers broader task-relevant regions, including the robot arm, table surface, and target objects, consistent with *tt* acting as a high-level semantic prior rather than a localized spatial tracker.

Temporal attention. As shown in the bottom row of Figure 7, *vt* and *tt* exhibit different temporal emphases rather than a simple object-tracking pattern. *vt* attention is more visible at the beginning and end of the sequence, suggesting that visual reasoning tokens are involved when the model anchors the initial state and resolves the final consequence of the motion. *tt* attention is less pronounced in the earliest frame, but becomes comparably salient across the later sampled frames, consistent with a prompt-level prior that remains active once the temporal context has been established. This contrast suggests a complementary temporal division: *vt* emphasizes boundary states in the visual evolution, while *tt* provides a steadier semantic constraint over the subsequent reasoning process.

5 Conclusion and Limitations

We present the OPENCOF-17K Dataset, a diverse video reasoning dataset for improving Chain-of-Frame reasoning in video generation models. Fine-tuning Wan2.2-I2V-A14B on the OPENCOF-17K Dataset yields WAN-COF, showing that diverse temporal supervision improves CoF behavior across four external benchmarks. We further explore Visual and Textual Reasoning Tokens and analyze how they organize reasoning state through benchmark results and attention patterns. One limitation is that *vt* and *tt* are investigated separately. How to effectively combine them remains future work.

Acknowledgments

This work is supported in part by Guangdong Basic and Applied Basic Research Foundation (No. 2023B1515130008, XW).

References

[1] Omer Bar-Tal, Hila Chefer, Omer Tov, Charles Herrmann, Roni Paiss, Shiran Zada, Ariel Ephrat, Junhwa Hur, Guanghui Liu, Amit Raj, et al. Lumiere: A space-time diffusion model for video generation. In *SIGGRAPH Asia 2024 Conference Papers*, pages 1–11, 2024.

[2] Zefan Cai, Haoyi Qiu, Tianyi Ma, Haozhe Zhao, Gengze Zhou, Kung-Hsiang Huang, Parisa Kordjamshidi, Minjia Zhang, Wen Xiao, Jiuxiang Gu, et al. Mmgr: Multi-modal generative reasoning. *arXiv preprint arXiv:2512.14691*, 2025.

[3] Harold Haodong Chen, Disen Lan, Wen-Jie Shu, Qingyang Liu, Zihan Wang, Sirui Chen, Wenkai Cheng, Kanghao Chen, Hongfei Zhang, Zixin Zhang, et al. Tivibench: Benchmarking think-in-video reasoning for video generative models. *arXiv preprint arXiv:2511.13704*, 2025.

[4] Xinyan Chen, Renrui Zhang, Dongzhi Jiang, Aojun Zhou, Shilin Yan, Weifeng Lin, and Hongsheng Li. Mint-cot: Enabling interleaved visual tokens in mathematical chain-of-thought reasoning. *arXiv preprint arXiv:2506.05331*, 2025.

[5] Hyung Won Chung, Noah Constant, Xavier Garcia, Adam Roberts, Yi Tay, Sharan Narang, and Orhan Firat. Unimax: Fairer and more effective language sampling for large-scale multilingual pretraining. *arXiv preprint arXiv:2304.09151*, 2023.

[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.

[7] 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.

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

[9] Google DeepMind. Veo-3 technical report. Technical report, Google DeepMind, May 2025.

[10] Ziyu Guo, Xinyan Chen, Renrui Zhang, Ruichuan An, Yu Qi, Dongzhi Jiang, Xiangtai Li, Manyuan Zhang, Hongsheng Li, and Pheng-Ann Heng. Are video models ready as zero-shot reasoners? an empirical study with the mme-cof benchmark. *arXiv preprint arXiv:2510.26802*, 2025.

[11] Ziyu Guo, Rain Liu, Xinyan Chen, and Pheng-Ann Heng. Atlas: Agentic or latent visual reasoning? one word is enough for both. *arXiv preprint arXiv:2605.15198*, 2026.

[12] Xuming He, Zehao Fan, Hengjia Li, Fan Zhuo, Hankun Xu, Senlin Cheng, Di Weng, Haifeng Liu, Can Ye, and Boxi Wu. Ruler-bench: Probing rule-based reasoning abilities of next-level video generation models for vision foundation intelligence. *arXiv preprint arXiv:2512.02622*, 2025.

[13] Jack Hong, Chenxiao Zhao, ChengLin Zhu, Weiheng Lu, Guohai Xu, and Xing Yu. Deepeyesv2: Toward agentic multimodal model. *arXiv preprint arXiv:2511.05271*, 2025.

[14] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Liang Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *Iclr*, 1(2):3, 2022.

[15] Yushi Hu, Weijia Shi, Xingyu Fu, Dan Roth, Mari Ostendorf, Luke Zettlemoyer, Noah A Smith, and Ranjay Krishna. Visual sketchpad: Sketching as a visual chain of thought for multimodal language models. *Advances in Neural Information Processing Systems*, 37:139348–139379, 2024.

[16] Dongzhi Jiang, Renrui Zhang, Ziyu Guo, Yanwei Li, Yu Qi, Xinyan Chen, Liuhui Wang, Jianhan Jin, Claire Guo, Shen Yan, et al. Mme-cot: Benchmarking chain-of-thought in large multimodal models for reasoning quality, robustness, and efficiency. *arXiv preprint arXiv:2502.09621*, 2025.

[17] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213, 2022.

[18] 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.

[19] Kuaishou Technology. Kling ai: Next-generation ai creative studio. https://klingai.com/, June 2024.

[20] Chengzu Li, Zanyi Wang, Jiaang Li, Yi Xu, Han Zhou, Huanyu Zhang, Ruichuan An, Dengyang Jiang, Zhaochong An, Ivan Vulić, et al. Thinking in frames: How visual context and test-time scaling empower video reasoning. *arXiv preprint arXiv:2601.21037*, 2026.

[21] 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.

[22] Yifan Li, Yukai Gu, Yingqian Min, Zikang Liu, Yifan Du, Kun Zhou, Min Yang, Wayne Xin Zhao, and Minghui Qiu. Viper: Process-aware evaluation for generative video reasoning. *arXiv preprint arXiv:2512.24952*, 2025.

[23] Xinxin Liu, Zhaopan Xu, Ming Li, Kai Wang, Yong Jae Lee, and Yuzhang Shang. Can world simulators reason? gen-vire: A generative visual reasoning benchmark. *arXiv preprint arXiv:2511.13853*, 2025.

[24] Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chun yue Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating math reasoning in visual contexts with gpt-4v, bard, and other large multimodal models. *ArXiv*, abs/2310.02255, 2023.

[25] Yang Luo, Xuanlei Zhao, Baijiong Lin, Lingting Zhu, Liyao Tang, Yuqi Liu, Ying-Cong Chen, Shengju Qian, Xin Wang, and Yang You. V-reasonbench: Toward unified reasoning benchmark suite for video generation models. *arXiv preprint arXiv:2511.16668*, 2025.

[26] Nanye Ma, Mark Goldstein, Michael S Albergo, Nicholas M Boffi, Eric Vanden-Eijnden, and Saining Xie. Sit: Exploring flow and diffusion-based generative models with scalable interpolant transformers. In *European Conference on Computer Vision*, pages 23–40. Springer, 2024.

[27] Fanqing Meng, Lingxiao Du, Zongkai Liu, Zhixiang Zhou, Quanfeng Lu, Daocheng Fu, Tiancheng Han, Botian Shi, Wenhai Wang, Junjun He, et al. Mm-eureka: Exploring the frontiers of multimodal reasoning with rule-based reinforcement learning. *arXiv preprint arXiv:2503.07365*, 2025.

[28] OpenAI. Openai o3 and o4-mini system card. Technical report, OpenAI, April 2025.

[29] OpenAI. Sora 2 system card. Technical report, OpenAI, September 2025.

[30] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4195–4205, 2023.

[31] Yu Qi, Xinyi Xu, Ziyu Guo, Siyuan Ma, Renrui Zhang, Xinyan Chen, Ruichuan An, Ruofan Xing, Jiayi Zhang, Haojie Huang, et al. Mme-cof-pro: Evaluating reasoning coherence in video generative models with text and visual hints. *arXiv preprint arXiv:2603.20194*, 2026.

[32] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[33] Team Seedance, De Chen, Liyang Chen, Xin Chen, Ying Chen, Zhuo Chen, Zhuowei Chen, Feng Cheng, Tianheng Cheng, Yufeng Cheng, et al. Seedance 2.0: Advancing video generation for world complexity. *arXiv preprint arXiv:2604.14148*, 2026.

[34] Team Seedance, Heyi Chen, Siyan Chen, Xin Chen, Yanfei Chen, Ying Chen, Zhuo Chen, Feng Cheng, Tianheng Cheng, Xinqi Cheng, et al. Seedance 1.5 pro: A native audio-visual joint generation foundation model. *arXiv preprint arXiv:2512.13507*, 2025.

[35] Hao Shao, Shengju Qian, Han Xiao, Guanglu Song, Zhuofan Zong, Letian Wang, Yu Liu, and Hongsheng Li. Visual cot: Advancing multi-modal language models with a comprehensive dataset and benchmark for chain-of-thought reasoning. *Advances in Neural Information Processing Systems*, 37:8612–8642, 2024.

[36] David Sharon and Nicole Brichtova. Image editing in gemini just got a major upgrade. The Keyword (Google Blog), August 26 2025.

[37] Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538*, 2017.

[38] Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, et al. Openai gpt-5 system card. *arXiv preprint arXiv:2601.03267*, 2025.

[39] Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.

[40] Jingqi Tong, Yurong Mou, Hangcheng Li, Mingzhe Li, Yongzhuo Yang, Ming Zhang, Qiguang Chen, Tianyi Liang, Xiaomeng Hu, Yining Zheng, et al. Thinking with video: Video generation as a promising multimodal reasoning paradigm. *arXiv preprint arXiv:2511.04570*, 2025.

[41] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[42] Homer Rich Walke, Kevin Black, Tony Z Zhao, Quan Vuong, Chongyi Zheng, Philippe Hansen-Estruch, Andre Wang He, Vivek Myers, Moo Jin Kim, Max Du, et al. Bridgedata v2: A dataset for robot learning at scale. In *Conference on Robot Learning*, pages 1723–1736. PMLR, 2023.

[43] Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, Jiayu Wang, Jingfeng Zhang, Jingren Zhou, Jinkai Wang, Jixuan Chen, Kai Zhu, Kang Zhao, Keyu Yan, Lianghua Huang, Mengyang Feng, Ningyi Zhang, Pandeng Li, Pingyu Wu, Ruihang Chu, Ruili Feng, Shiwei Zhang, Siyang Sun, Tao Fang, Tianxing Wang, Tianyi Gui, Tingyu Weng, Tong Shen, Wei Lin, Wei Wang, Wei Wang, Wenmeng Zhou, Wente Wang, Wenting Shen, Wenyuan Yu, Xianzhong Shi, Xiaoming Huang, Xin Xu, Yan Kou, Yangyu Lv, Yifei Li, Yijing Liu, Yiming Wang, Yingya Zhang, Yitong Huang, Yong Li, You Wu, Yu Liu, Yulin Pan, Yun Zheng, Yuntao Hong, Yupeng Shi, Yutong Feng, Zeyinzi Jiang, Zhen Han, Zhi-Fan Wu, and Ziyu Liu. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

[44] Maijunxian Wang, Ruisi Wang, Juyi Lin, Ran Ji, Thaddäus Wiedemer, Qingying Gao, Dezhi Luo, Yaoyao Qian, Lianyu Huang, Zelong Hong, et al. A very big video reasoning suite. *arXiv preprint arXiv:2602.20159*, 2026.

[45] 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.

[46] 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.

[47] Bing Wu, Chang Zou, Changlin Li, Duojun Huang, Fang Yang, Hao Tan, Jack Peng, Jianbing Wu, Jiangfeng Xiong, Jie Jiang, et al. Hunyuanvideo 1.5 technical report. *arXiv preprint arXiv:2511.18870*, 2025.

[48] Yi Xu, Chengzu Li, Han Zhou, Xingchen Wan, Caiqi Zhang, Anna Korhonen, and Ivan Vulić. Visual planning: Let's think only with images. *arXiv preprint arXiv:2505.11409*, 2025.

[49] Cheng Yang, Haiyuan Wan, Yiran Peng, Xin Cheng, Zhaoyang Yu, Jiayi Zhang, Junchi Yu, Xinlei Yu, Xiawu Zheng, Dongzhan Zhou, et al. Reasoning via video: The first evaluation of video models' reasoning abilities through maze-solving tasks. *arXiv preprint arXiv:2511.15065*, 2025.

[50] Yi Yang, Xiaoxuan He, Hongkun Pan, Xiyan Jiang, Yan Deng, Xingtao Yang, Haoyu Lu, Dacheng Yin, Fengyun Rao, Minfeng Zhu, et al. R1-onevision: Advancing generalized multimodal reasoning through cross-modal formalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2376–2385, 2025.

[51] Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou, Pan Lu, Kai-Wei Chang, Peng Gao, et al. Mathverse: Does your multi-modal llm truly see the diagrams in visual math problems? *ECCV 2024*, 2024.

[52] Ziwei Zheng, Michael Yang, Jack Hong, Chenxiao Zhao, Guohai Xu, Le Yang, Chao Shen, and Xing Yu. Deepeyes: Incentivizing" thinking with images" via reinforcement learning. *arXiv preprint arXiv:2505.14362*, 2025.

A Appendix

Overview

We organize our supplementary material as follows.

- Related Work
  - Modern Video Generation Models
  - Video Reasoning Benchmarks
  - Video Generation Models for Reasoning
- Dataset Details
  - Dataset Examples
  - Dataset Statistics
- Additional Implementation Details
  - Backbone and LoRA Setup
  - Model Variants
  - Compared Systems
  - Evaluation Protocol
  - Compute Resources
- Additional Ablation Study
  - Data Composition and Diversity
  - Task-level ID/OOD Analysis
  - Reasoning-token Interventions and Joint Model

A.1 Related Work

A.1.1 Modern Video Generation Models

Video generation models have advanced rapidly in visual quality, temporal consistency, and instruction following. Architecturally, the shift from U-Net-based diffusion [1, 32] to DiT [30] and flow-matching frameworks [26] has enabled scalable training and higher-resolution synthesis. Open systems such as HunyuanVideo [18], HunyuanVideo-1.5 [47], and Wan [43] provide strong backbones for image-to-video and text-to-video generation, while proprietary systems including Kling [19], Veo [9], Seedance [33], and Sora [29] continue to push generation quality and controllability. These advances make video generation a promising substrate for visual reasoning, but high-fidelity generation alone does not guarantee coherent multi-step reasoning.

A.1.2 Video Reasoning Benchmarks

Reasoning has become a central capability for multimodal models [16, 4], but conventional chain-of-thought pipelines [45] often rely on text, static visual observations, external tools [15], or auxiliary image generation [21]. Video generation provides another path, where reasoning unfolds through temporally connected frames [46]. Recent benchmarks, including MME-CoF [10], VideoThinkBench [40], V-ReasonBench [25], Gen-ViRe [23], RULER-Bench [12], MMGR [2], TiViBench [3], and VIPER [22], formalize and evaluate this capability from different perspectives. OpenCoF builds on these benchmarks but focuses on improving CoF reasoning through the OpenCoF-17K dataset and reasoning-token exploration.

A.1.3 Video Generation Models for Reasoning

Training video generation models for reasoning remains less explored than benchmarking. VR-Bench [49] studies maze-solving and test-time scaling, Thinking in Frames [20] focuses on domains such as mazes and tangram puzzles, and VBVR [44] scales video-reasoning data within a cognitive task suite. These efforts suggest that targeted supervision improves video reasoning, but they are often domain-specific or internally coupled between training and evaluation. OpenCoF-17K instead combines four data sources and supports transfer evaluation on independent external benchmarks, while OpenCoF further explores how dedicated reasoning techniques organize reasoning state through Visual and Textual Reasoning Tokens.

Table 6: Per-benchmark evaluation protocol. "Judge" denotes the scoring model used to evaluate each generated video.

Benchmark	Headline metric	Judge model	Averaging
MME-CoF [10]	Overall (0–4)	Gemini-2.5-Pro	3 judge passes per video
Gen-ViRe [23]	Avg. normalized score	Gemini-2.5-Pro	5 generations per prompt
VIPER [22]	POC@1.0	GPT-5	single pass
RULER-Bench [12]	Overall (across IF/VC/VF/RC)	GPT-o3	single pass

A.2 Dataset Details

Dataset examples. Figures 8 and 9 show representative cases for each of the 11 task families in the OpenCoF-17K dataset. For every case we display three uniformly sampled frames from the demonstration video together with the original text prompt. The selected examples cover all four curation pipelines described in Section 2.2 and span structured-board reasoning (Chess, Sudoku), expert-guided geometric reasoning (2D Geometry, Dot-to-dot), procedural spatial and physical reasoning (Tangram Puzzle, Cube Folding, 3D Polycube Rotation, Physics Motion), and reused real-world demonstrations (Maze, Embodied Manipulation, VBVR).

A.3 Additional Implementation Details

Backbone and LoRA setup. Wan2.2-I2V-A14B adopts a Mixture-of-Experts [37] denoising architecture with separate high-noise and low-noise experts across diffusion timesteps, an umT5 [5] text encoder, a Wan-VAE for visual encoding and decoding, and a mainstream DiT denoiser. We fine-tune all three models (Wan-CoF, Wan-CoF$_{vt}$, and Wan-CoF$_{tt}$) from the same Wan2.2-I2V-A14B backbone using LoRA [14], applied only to the DiT denoiser while the umT5 text encoder and the Wan-VAE are kept frozen. Within each DiT block, LoRA targets the attention projections (`q`, `k`, `v`, `o`) together with the two MLP linear layers (`ffn.0`, `ffn.2`), with rank set to 32. The learning rate is configured to 2×10^{-5}.

Model variants. To attribute every gain to the dataset alone, we keep this architecture entirely unchanged and fine-tune it on OpenCoF-17K with LoRA, yielding Wan-CoF. For reasoning-token exploration, we run two independent fine-tuning experiments with LoRA on Wan2.2-I2V-A14B, one with only *vt* and one with only *tt*, yielding the models Wan-CoF$_{vt}$ and Wan-CoF$_{tt}$, where we set $N_r = N_t = 16$. Wan-CoF is trained on OpenCoF-17K for 2 epochs, while Wan-CoF$_{vt}$ and Wan-CoF$_{tt}$ are trained for 5 epochs to give the additional reasoning-token parameters more updates to converge.

Compared systems. We compare Wan-CoF with both closed-source and open-source video generation models. The closed-source set covers state-of-the-art systems including Kling-v1 [19], the Seedance series [33] Seedance-1.0-Pro and Seedance-1.5, the Veo-3 [9] series Veo-3.0-Fast, Veo-3.0-Preview, Veo-3.0 and Veo-3.1, and Sora-2 [29]. The open-source set includes HunyuanVideo-I2V [18] and HunyuanVideo-1.5 [47].

Evaluation protocol. We evaluate our models on four video reasoning benchmarks: MME-CoF [10], RULER-Bench [12], VIPER [22], and Gen-ViRe [23]. Together, these benchmarks span reasoning categories well beyond the 11 task types in OpenCoF-17K. For RULER-Bench specifically, we restrict evaluation to the image-to-video cases, since our baseline does not support text-to-video generation. For every benchmark we follow its official judge model, as summarized in Table 6. Specifically, MME-CoF runs Gemini-2.5-Pro [6] on each generated video three times and reports the averaged score; Gen-ViRe samples five videos per prompt and averages the Gemini-2.5-Pro scores over them, as its official setting; VIPER reports POC@1.0 judged by GPT-5 [38] in a single pass; and RULER-Bench evaluates each video once with GPT-o3 [28] over its IF/VC/VF/RC checklist questions. We do not introduce any additional repetition or judge model on top of these defaults.

Compute resources. All training and inference are performed on an internal cluster of high-memory NVIDIA GPUs. Each fine-tuning run uses multi-GPU data parallelism and completes within a small

Chess



Please play as White or Black (determine your side by looking at the board and piece positions) and deliver checkmate on your next move.

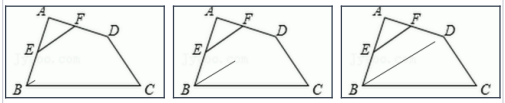
Maintain a fixed camera angle and lighting throughout the game. Static scene, no zoom, no pan, no dolly.

Sudoku



Create a static, smooth, animation that solves the given 4x4 sudoku. Enter the missing numbers one by one. Do not change anything else in the picture. Only fill the numbers in the empty cells so the sudoku is solved properly. A cursor moves and fills the correct number in the empty boxes.

2D Geometry



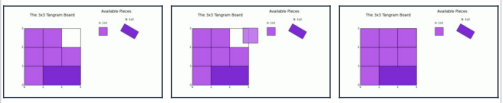
Smoothly connecting point B and point D with a line. The video ends once the connection process is complete. Static view, no zoom or pan.

Dot-to-dot



Animate the dots connecting sequentially from 1 to 12, each straight line appearing smoothly until the full outline emerges. Keep the background with the smiling sun and plants unchanged. Static shot.

Tangram Puzzle



Place piece A at (x, y) = (2, 2).

Cube Folding



Fold the net upward into a cube, keeping the camera fixed.

3D Polycube Rotation



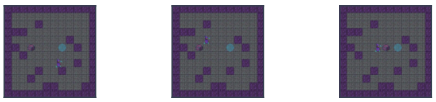
A structure composed of 10 cubes viewed from 318° azimuth, 32° elevation. The camera executes a smooth horizontal orbit around the structure, transitioning to a 95° azimuth which is a 137-degree rotation while maintaining a constant distance from the object.

Physics Motion



A ball is placed at the top of an inclined plane (inclination angle θ = 21.2 deg). The surface has a friction coefficient μ = 0.15. The ball is released and rolls down the slope onto the horizontal ground. Show the complete motion. Static shot.

Maze



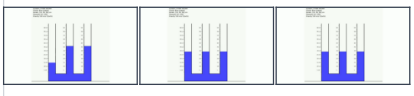
Create a 2D animation based on the provided image of a grid puzzle. The purple wizard moves into position behind the purple wooden crate and smoothly pushes it toward the blue magic circle. The purple wooden crate only slides when pushed from behind by the purple wizard and moves in a straight line along the gray stone bricks tiles. …

Embodied Manipulation



fold the cloth from right to left

VBVR



A system of 3 communicating vessels with equal-diameter vertical tubes is filled with water (water-like (low viscosity)), which appears blue in color. As shown in the initial frame, the liquid levels in the tubes are [14, 36, 36] cm respectively. Due to pressure differences between the tubes, the liquid begins to flow through the connecting channels at the bottom. The flow is governed by hydrostatic pressure equalization and damped by viscous resistance with coefficient k=3.92. As the liquid redistributes, the height differences gradually decrease, and the system evolves toward equilibrium. Eventually, through conservation of volume, all tubes reach the same final liquid level, which equals the average of the initial heights. Simulate this settling process from the initial unbalanced state to the final stable equilibrium.

Figure 8: Representative examples from OpenCoF-17K.

Chess



Please play as White or Black (determine your side by looking at the board and piece positions) and deliver checkmate on your next move.

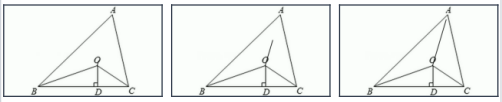
Maintain a fixed camera angle and lighting throughout the game. Static scene, no zoom, no pan, no dolly.

Sudoku



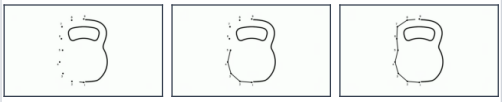
Create a static, smooth, animation that solves the given 4x4 sudoku. Enter the missing numbers one by one. Do not change anything else in the picture. Only fill the numbers in the empty cells so the sudoku is solved properly. A cursor moves and fills the correct number in the empty boxes.

2D Geometry



Smoothly connecting point O and point E with a line. The video ends once the connection process is complete. Static view, no zoom or pan.

Dot-to-dot



Animate the dots connecting sequentially from 1 to 9, each straight line appearing smoothly until the full outline emerges. Keep the background with the smiling sun and plants unchanged. Static shot.

Tangram Puzzle



Rotate piece A by 60 degrees clockwise, and place it at (x, y) = (2, 0).

Cube Folding



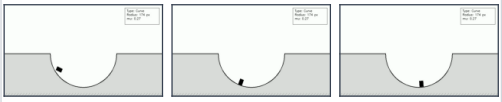
The net is folded to form a single cube, with folding edges clearly shown. Static camera perspective, no zoom or pan.

3D Polycube Rotation



Rotating around the 3D shape to suggest spatial structure, keeping lighting and geometry fixed.

Physics Motion



A block is released from the top edge of a U-shaped semicircular track (radius R = 174). The surface has a friction coefficient $\mu = 0.27$. The block slides down along the curved surface, passes through the lowest point, and climbs up the opposite side. Show the complete motion. Static shot.

Maze



Create a 2D animation based on the provided image of a maze. The blue circle slides smoothly along the white square path, stopping perfectly on the green circle. The blue circle never slides into or crosses the red x (trap areas). …

Embodied Manipulation



move the fork above cloth

VBVR



A black ball starts with the shown initial direction arrow. It moves in straight lines and reflects off the two vertical walls following the law of reflection (the angle of reflection equals the angle of incidence). Follow this unique trajectory until the ball hits one of the red hollow circle targets in the lower-right area. Identify which target circle is hit and show the full trajectory step by step.

Figure 9: Representative examples from OpenCoF-17K.

Table 7: **Data composition and diversity ablations.** The 7,600-example rows compare different source and task-family coverage at a matched scale; the 17,312-example rows compare task-family coverage at the same nominal data volume and optimization budget.

Training data	# Samples	Families	MME-CoF	RULER-Bench	VIPER	Gen-ViRe
Wan2.2-I2V-A14B (Baseline)	0	–	1.000	55.80	3.30	0.304
VBVR-only	7,750	1 (30 subtasks)	1.160	56.30	5.30	0.347
Newly-constructed-only	7,562	8	1.210	**56.40**	5.10	0.357
OpenCoF-17K downsample	7,600	11	**1.210**	56.00	**5.96**	**0.359**
Random 5-family mixture A	17,312	5	1.187	55.96	6.18	0.349
Random 5-family mixture B	17,312	5	1.163	55.82	5.74	0.354
Random 5-family mixture C	17,312	5	1.142	55.71	5.46	0.338
5-family mean ± s.d.	17,312	5	1.164±0.023	55.83±0.13	5.79±0.36	0.347±0.008
Wan-CoF (full OpenCoF-17K)	17,312	11	**1.300**	**56.80**	**7.50**	**0.391**

Table 8: **Task-level ID/OOD decomposition.** Scores are re-aggregated over task types that overlap with (ID) or are unseen in (OOD) OpenCoF-17K.

Model	MME-CoF		Gen-ViRe	
	ID (7)	OOD (5)	ID (14)	OOD (10)
Wan2.2-I2V-A14B (Baseline)	1.023	0.968	0.284	0.334
Wan-CoF	**1.417**	**1.128**	**0.365**	**0.428**
Δ *over the Baseline*	+0.394	+0.160	+0.081	+0.094

number of days, depending on the configured number of training epochs; generating videos for the four benchmark suites consumes comparable compute on the same cluster.

A.4 Additional Ablation Study

A.4.1 Data Composition and Diversity

To disentangle the effect of broader task coverage from raw sample count, we design two complementary matched controls. For the equal-scale subset comparison, we train on all 7,750 VBVR examples, all 7,562 newly constructed examples from pipelines 1–3, or a 7,600-example downsample of the full 11-family mixture. For the equal-volume diversity comparison, we independently sample three sets of five task families and construct a 17,312-example mixture for each while preserving the selected families' relative proportions. All variants use the same backbone, LoRA configuration, hyperparameters, and two-epoch training schedule as Wan-CoF; the results are reported in Table 7.

At the 7,600-example scale, the full-coverage downsample matches or exceeds the VBVR-only and newly-constructed-only variants on three of four benchmarks. At the same 17,312-example budget, the full 11-family mixture outperforms all three random 5-family mixtures on every benchmark. These matched controls show that the full-mixture advantage is not explained by raw sample count alone and that broader task-family coverage contributes measurably to CoF performance.

A.4.2 Task-level ID/OOD Analysis

To determine whether the observed transfer gains are confined to benchmark tasks that overlap with our training families, we decompose MME-CoF and Gen-ViRe at the task level. Each benchmark task is compared with the OpenCoF-17K task inventory and capability descriptions and manually verified; the split is fixed before inspecting model improvements. Table 8 reports the re-aggregated scores.

Wan-CoF improves on both subsets of both benchmarks. MME-CoF improves more on ID tasks ($+0.394$ versus $+0.160$), whereas Gen-ViRe shows a slightly larger gain on OOD tasks ($+0.094$ versus $+0.081$). The gains are therefore not confined to task types that overlap with OpenCoF-17K.

A.4.3 Additional Reasoning-token Analyses

Token removal and randomization. To test whether the gains arise from the learned token content rather than merely from adding extra token slots, we perform two causal interventions on the trained

Table 9: **Reasoning-token interventions on MME-CoF.** *None* removes the token pathway; *random* replaces only the learned token values while retaining the trained checkpoint.

Variant	Token state	Overall	Instruction Alignment	Temporal Consistency	Visual Stability	Content Fidelity	Focus Relevance
vt	trained	**1.34**	**0.26**	**1.38**	**2.33**	**0.99**	**1.76**
vt	none	1.26	0.24	1.21	2.16	0.94	1.74
vt	random	1.27	0.23	1.19	2.21	0.96	1.74
tt	trained	**1.35**	**0.34**	**1.27**	**2.18**	**1.17**	**1.77**
tt	none	1.28	0.28	1.20	2.10	1.08	1.76
tt	random	1.25	0.25	1.17	2.08	1.02	1.73

Table 10: **Joint reasoning-token ablation.**

Model	MME-CoF	Gen-ViRe	RULER-Bench	VIPER
WAN-COF	1.30	0.391	56.8	7.5
WAN-COF$_{vt}$	1.34	**0.441**	**59.6**	8.2
WAN-COF$_{tt}$	**1.35**	0.406	58.4	**8.8**
WAN-COF-COMB	**1.35**	0.412	56.9	7.3

WAN-COF$_{vt}$ and WAN-COF$_{tt}$ checkpoints. We either remove the tokens (*none*) or re-draw only their values from the original initializer (*random*), which preserves parameter count, sequence length, and inference computation. Both interventions reduce every MME-CoF dimension in Table 9. The mean relative decrease is 12.9% over Instruction Alignment and Temporal Consistency, versus 4.5% over Visual Stability, Content Fidelity, and Focus Relevance, showing that performance depends on the learned token content rather than merely on additional token slots.

Joint visual and textual reasoning tokens. To examine whether the complementary visual and textual pathways can combine their individual strengths, we train WAN-COF-COMB with *vt* and *tt* together under the single-token settings. It matches the best single-token result on MME-CoF but trails the best result on the other benchmarks, suggesting interference between the two conditioning pathways (Table 10).

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The three main conclusions and the scope of our paper are detailed in the final section of the introduction and summarized in the abstract.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss the limitations of the work in Conclusion.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: All the assumptions and theorems are clearly stated.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The architecture details are provided in the Experiments and Exploration sections for reproducing the main experimental results.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: The code and data are released on our project pag.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: The training and test details are provided in the Experiments section.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We draw error bars in the attention-pattern figure for reasoning tokens. Moreover, Gen-ViRe uses five generations per prompt.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.

- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The compute resources are provided in the appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: Our research conforms with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: There is no societal impact of the work performed.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper poses no such risks.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We cite the original papers or websites that produced the code package or dataset.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: We detail the data and model in the experiments section and appendix.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: This paper does not involve crowdsourcing or research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: This paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [Yes]

Justification: We employ Gemini-2.5-Pro, GPT-5 and GPT-o3 as automated evaluators (LLM-as-a-judge) to score the experimental results.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.