

DeformM: Reasoning-Guided Physics-Aware Video Generation via Spatial-Temporal Masking

Yunyi Li, Yu Qiao, Yaohui Wang, and Xinyuan Chen*

Shanghai Artificial Intelligence Laboratory, Shanghai, China



Fig. 1: DeformM is an image-to-video physics-aware generation framework. It generates realistic deformation dynamics by leveraging VLM-guided physical reasoning to focus on physics-critical regions.

Abstract. Video generation models achieve high visual quality but often struggle to generate physics-aware videos. Unlike rigid-body motion, which can be described by explicit trajectories or formulas, complex deformation dynamics remain challenging to synthesize. We observe that a lack of physical reasoning for localizing dynamic areas allows irrelevant regions to dilute the model’s attention, leading to generation failure. In this paper, we propose **DeformM**, a reasoning-guided image-to-video generation framework that directs the model’s focus toward physics-critical regions. To reason and localize these critical regions, we introduce a VLM-guided physical reasoning module, DeformM-Reason, to identify target objects and generate spatial-temporal masks. For physical guidance, we develop two alternative strategies: DeformM-Free for training-free mechanism analysis, and DeformM-Injection as a powerful training-based generator. Experimental results demonstrate that DeformM improves the realism of generated deformation scenarios, outperforming baseline models in both visual quality and physical consistency.

* Corresponding author.

Keywords: Video generation · Physics-aware video generation

1 Introduction

Video generation models have achieved significant advances, attracting growing attention for film production [14, 23, 26, 31] and world modeling [9, 25]. While such models provide realistic visuals for simulating virtual environments, these models do not guarantee underlying consistency with the laws of physics [18, 20, 30]. Beyond improving visual quality, ensuring physical consistency is essential yet remains an unresolved challenge for practical video generators.

Despite various efforts [15, 30, 34] to address this issue, existing approaches still face significant bottlenecks. Specifically, simulator-based methods [16, 18, 20] require explicit scene reconstruction, which limits their scalability to diverse real-world environments. Alternatively, data-driven and reward-based methods [3, 24] are often constrained by limited physical data diversity and complex reward engineering. More recently, some approaches utilize visual large language models (VLMs) [2, 28, 29] to steer video generation towards physical plausibility [30, 35], but are restricted to rigid-body dynamics. Unlike rigid motion, deformation dynamics, such as melting, squeezing, and fracturing, involve high-dimensional transformations that are difficult to characterize through simple formulas or trajectory sequences. Furthermore, existing generation models struggle to reason about the physical properties of objects based on the visual and textual inputs, thus failing to focus on physics-critical regions and leading to physically implausible artifacts.

To this end, we propose **DeformM**, a reasoning-guided image-to-video generation framework that directs the video generation model’s focus toward physics-critical regions to generate realistic deformation dynamics. To achieve this, two crucial questions must be addressed: *how to leverage physical reasoning to precisely localize the regions where deformation dynamics occur* and *how to effectively steer the model’s attention toward these critical areas*. Figure 1 illustrates the motivation behind DeformM and shows representative results of our method.

VLM-guided Physical Reasoning. To empower the video generation process with physical property reasoning and precise phenomena localization, we propose a VLM-guided physical reasoning module, called **DeformM-Reason**. This module acts as a cognitive agent by identifying the target objects most relevant to the deformation described in the text prompt and the initial image. We then employ an open-vocabulary object detector to ground these objects and generate a precise spatial-temporal mask. The mask is then aligned with video latents, providing concrete geometric priors to inform the model where to apply physical laws throughout the generation process.

Spatial-temporal Mask Injection. To integrate physical priors into the generation process, we provide two alternative strategies for mask-guided modulation: **DeformM-Free** and **DeformM-Injection**. We initiate our study with DeformM-Free, a training-free approach that modulates cross-attention queries [22] to concentrate generative capacity on the identified regions. This plug-and-play

exploration reveals a fundamental observation: achieving physical consistency primarily requires the model to refocus its globally dispersed attention onto physics-critical areas. Building upon this insight, we further develop DeformM-Injection as a more powerful, training-based alternative to push the boundaries of deformation realism. It employs a dense mask modulator to inject high-dimensional features into both cross-attention and self-attention blocks, optimized via a specialized region-sensitive loss. While both methods independently enable physics-aware video generation, DeformM-Free unveils the underlying mechanism through step-by-step analysis, whereas DeformM-Injection provides a powerful generator for diverse and realistic deformation.

We summarize our contributions as follows: 1) We propose DeformM, a reasoning-guided image-to-video generation framework to generate realistic deformation dynamics by guiding the model’s focus toward essential physical regions. 2) We introduce a VLM-guided physical reasoning module, DeformM-Reason, for precise spatial-temporal localization of deformation. Furthermore, we develop two alternative methods: DeformM-Free, which facilitates the analysis and validation of our core observations, and DeformM-Injection, a powerful training-based generator designed for diverse and realistic deformation dynamics. 3) Extensive experiments show that DeformM surpasses state-of-the-art (SOTA) methods in both visual quality and physical consistency, achieving physics-aware video generation.

2 Related Work

2.1 Physics-Aware Video Generation

Achieving physics-aware video generation is an essential objective for practical applications, spurring numerous studies [18, 24, 30]. These approaches can be categorized by the source of physical information they incorporate. Some works, like PhysGen [18], PhysMotion [20], and WonderPlay [16], rely on simulator-based frameworks to generate videos that adhere to physical laws. However, the requirement for building complex assets and the inherent complexity of their extended pipelines limit their efficiency. Another branch of research focuses on learning physical laws from large-scale data. For instance, WISA [24], PISA Experiments [15], and Force Prompting [8] attempt to acquire physical priors from a vast amount of real-world or simulated data. Furthermore, methods represented by PhysRVG [34] and PhysHPO [3] employ reinforcement learning to capture physical information. However, these methods rely heavily on the design of sophisticated reward functions and are prone to reward hacking. Lastly, a burgeoning category of methods leverages VLMs to guide video generation models in producing physically plausible videos. Specifically, VLIPP [30] and PhyRPR [35] utilize VLMs to plan object trajectories for training-free video generation. However, relying on trajectory bounding boxes or simple formulas fails to capture fine-grained deformation, inherently limiting these approaches to rigid-body motions, such as collisions and free falls. Furthermore, current generation models primarily focus on visual pattern synthesis, lacking the explicit

physical reasoning capabilities required to understand the physical dynamics of complex scenes.

2.2 VLM-Guided Visual Generation

VLMs exhibit robust multimodal understanding and interaction capabilities, which are leveraged to enhance the quality of visual generation and editing. Some studies synergize the generative power of diffusion models with the advanced reasoning and planning capabilities of VLMs. For image editing, R-Genie [33] utilizes VLM-based reasoning to address the lack of deep comprehension regarding user intentions and contextual reasoning. DiffSensei [27] employs VLMs to decompose complex prompts into structured goals, ensuring precise alignment with multi-objective instructions. For video generation, VLIPP [30] and PhyRPR [35] utilize VLMs to plan plausible trajectories for object motion. Alternatively, several approaches employ VLMs as evaluators to improve generation quality through closed-loop feedback. GenPilot [32] utilizes VLM-based visual question answering and image captioning for systematic error analysis, guiding iterative prompt optimization. Similarly, VIVA [5] introduces the Edit-GRPO algorithm for video editing, which leverages VLMs to provide relative rewards based on instruction fidelity, content preservation, and aesthetic quality.

3 Deform Framework

DeformM aims to generate realistic deformation videos that adhere to physical dynamics, covering melting, squeezing, fracturing, stretching, and slicing. We focus on the image-to-video (I2V) task of synthesizing the video $\mathbf{x}_{1:T}$ from an initial image c_{img} and a textual prompt c_{text} , represented as:

$$(c_{img}, c_{text}) \xrightarrow{\text{DeformM}} \mathbf{x}_{1:T}, \quad (1)$$

where T denotes the number of generated frames. Our primary objective is to ensure that the output video $\mathbf{x}_{1:T}$ is physically plausible, accurately capturing intricate physical evolutions described in c_{text} .

As illustrated in Fig. 2, DeformM comprises a reasoning module, DeformM-Reason, and two alternative strategies for the model’s attention redirection: DeformM-Free and DeformM-Injection. Specifically, DeformM-Reason integrates VLM-guided physical reasoning to analyze the input text prompt c_{text} and the initial image c_{img} for precise localization of physics-critical regions (Sec. 3.2). To redirect the model’s attention, DeformM-Free employs a training-free mechanism to incorporate physical priors into the inference process (Sec. 3.3). This preliminary exploration reveals the significance of explicit reasoning and attention anchoring. Building upon these insights, the training-based DeformM-Injection is proposed to enhance deformation realism and scene diversity, synthesizing physics-aware videos (Sec. 3.4).

3.1 Preliminary: Flow Matching for Video Generation

Flow matching has emerged as a powerful paradigm for high-fidelity video generation due to its superior efficiency and sampling quality [6, 13, 17, 23]. In this work, we build our Deform framework upon a pre-trained flow-matching-based video generator. In the training stage, video flow matching models first map an input video $\mathbf{x}$ into a latent tensor $\mathbf{z}_0 = \mathcal{E}(\mathbf{x})$ via a pretrained encoder $\mathcal{E}(\cdot)$ [19]. Instead of predicting noise, the model $\mathbf{v}_\theta$ learns to predict the velocity field governing the probability path between the data distribution and the noise distribution. We adopt the Optimal Transport (OT) path [1, 17], where the intermediate state $\mathbf{z}_t$ at timestep $t \in [0, 1]$ is constructed as:

$$\mathbf{z}_t = (1 - t)\mathbf{z}_0 + t\mathbf{z}_1, \quad (2)$$

where $\mathbf{z}_1 \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ represents Gaussian noise. The model is trained to regress the ground-truth velocity $\mathbf{u}_t(\mathbf{z}_t | \mathbf{z}_0, \mathbf{z}_1) = \mathbf{z}_1 - \mathbf{z}_0$ by minimizing:

$$\mathcal{L}(\theta) = \mathbb{E}_{\mathbf{z}_0, \mathbf{z}_1, t, c} \|\mathbf{v}_\theta(\mathbf{z}_t, t, c) - (\mathbf{z}_1 - \mathbf{z}_0)\|^2, \quad (3)$$

where c denotes conditions [10], such as text and image prompts, and t is sampled uniformly from $[0, 1]$. In the inference stage, video generation is formulated as solving an Ordinary Differential Equation (ODE) [4]. Starting from a noise sample $\mathbf{z}_1 \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, the model generates the latent video $\mathbf{z}_0$ by numerically integrating the predicted velocity field $\mathbf{v}_\theta$ from $t = 1$ to $t = 0$.

3.2 Deform-Reason: VLM-Guided Physical Reasoning

A prerequisite for physics-aware video generation is the precise identification of where the physical laws should be applied. However, existing video generation models often lack such physical reasoning capabilities. Specifically, while a text prompt c_{text} may specify deformation dynamics, the model often struggles to autonomously associate this semantic instruction with the corresponding pixels in the input image c_{img} . We observe that this semantic-spatial gap is a primary bottleneck. To bridge this gap, we propose Deform-Reason, a VLM-guided physical reasoning module that acts as a cognitive agent, translating abstract physical instructions into concrete spatial-temporal constraints.

As illustrated in **Algorithm 1**, given c_{img} and c_{text} , we first employ a VLM for physical reasoning to identify the object most relevant to the described deformation, denoted as $\mathcal{R}(c_{img}, c_{text}) \rightarrow \mathcal{O}_{target}$, where $\mathcal{O}_{target}$ is the semantic identifier of the target object. Subsequently, an open-vocabulary object detector is utilized to establish spatial grounding. The grounding process varies between the training and inference stages:

- **Training Stage.** We utilize the ground-truth video sequences to obtain precise bounding box sequences. The detector identifies the target object $\mathcal{O}_{target}$ across all frames $i \in \{1, \dots, T\}$, yielding a sequence of bounding boxes $\{\mathbf{B}_i\}_{i=1}^T$.

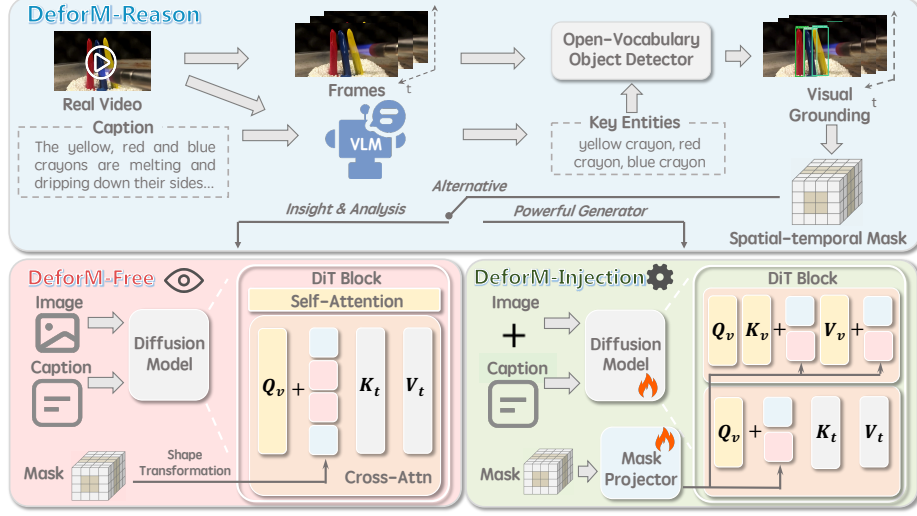


Fig. 2: Illustration of DeformM. DeformM takes a text prompt and an initial image as inputs. The DeformM-Reason module first generates spatial-temporal masks by localizing physics-critical regions via VLM-guided reasoning. For physical guidance, we provide two alternative strategies: (1) DeformM-Free, a training-free method for mechanism analysis and observation validation. (2) DeformM-Injection, a powerful training-based generator for diverse and realistic deformation dynamics.

- **Inference Stage.** Since the future frames are unavailable during inference, we leverage the VLM to reason about the temporal localization of the object. The detector first locates the object in c_{img} to obtain $\mathbf{B}_1$. The VLM then performs temporal planning to extrapolate the first box into a sequence of spatial-temporal constraints $\{\mathbf{B}_i\}_{i=2}^T$, accounting for the expected deformation described in c_{text} .

Finally, we construct a binary spatial-temporal mask $\mathbf{M} \in \{0, 1\}^{T \times H \times W}$ by setting the interior of each $\mathbf{B}_i$ to 1. To align with the generative process, $\mathbf{M}$ is downsampled to the dimensions of the video latents $\mathbf{z} \in \mathbb{R}^{T' \times H' \times W'}$, where T', H', W' denote the temporal and spatial dimensions in the latent space. The resulting mask serves as an explicit spatial-temporal prior, effectively guiding the model to concentrate its generative capacity on regions where deformation dynamics occur.

By transforming the implicit physical potential of an image into an explicit spatial-temporal mask, we provide the foundation for the subsequent training-free attention modulation and training-based methods, effectively guiding the model to focus on the regions where physics matters.

Algorithm 1 DeformM-Reason: VLM-Guided Physical Reasoning

Input: Image c_{img} , prompt c_{text} , video length T , mode $\in \{\text{train}, \text{infer}\}$.
Output: Latent-aligned mask $\mathbf{M}$.

- 1: $\mathcal{O}_{target} \leftarrow \text{REASON}(c_{img}, c_{text})$ {Identify target via VLM}
- 2: **if** mode = train **then**
- 3: $\{\mathbf{B}_i\}_{i=1}^T \leftarrow \text{DETECT}(\mathcal{O}_{target}, \text{GT frames})$
- 4: **else**
- 5: $\mathbf{B}_1 \leftarrow \text{DETECT}(\mathcal{O}_{target}, c_{img})$
- 6: $\{\mathbf{B}_i\}_{i=2}^T \leftarrow \text{PLAN}(\mathbf{B}_1, c_{text})$ {VLM-based temporal extrapolation}
- 7: **end if**
- 8: $\mathbf{M}_{mask} \leftarrow \mathbf{0} \in \{0, 1\}^{T \times H \times W}$
- 9: **for** $i = 1$ **to** T **do**
- 10: $\mathbf{M}_{mask}[i, (x, y) \in \mathbf{B}_i] \leftarrow 1$
- 11: **end for**
- 12: $\mathbf{M} \leftarrow \text{DOWNSAMPLE}(\mathbf{M}_{mask})$ {Align with latent shape (T', H', W') }
- 13: **return** $\mathbf{M}$

3.3 DeformM-Free: Spatial-Temporal Attention Modulation

The primary challenge in physics-aware video generation lies in the discrepancy between linguistic physical priors and localized visual generation. In this section, our empirical analysis reveals that the failure of base models to generate realistic deformation stems from this semantic-spatial gap. As illustrated in Fig. 3, the base model fails to concentrate the *melting* instruction on the target object regions, resulting in physically inconsistent generation where the melting occurs in irrelevant areas.

By visualizing the cross-attention maps [22] for the keyword *melting*, we observe that the attention signals are diluted across the entire frame. This observation confirms that the model struggles to autonomously associate abstract physical keywords with the corresponding spatial pixels. This insight serves as the foundation for our proposed DeformM-Free, as a training-free modulation method designed to bridge this gap by explicitly anchoring the model’s attention.

We leverage the spatial-temporal mask $\mathbf{M} \in \{0, 1\}^{T' \times H' \times W'}$ obtained from DeformM-Reason as a prior to modulate the attention mechanism. Let $\mathbf{z}$ be the video latent features and $\mathbf{e}_{text}$ be the text prompt embedding. In the cross-attention blocks, the queries $\mathbf{Q}$, keys $\mathbf{K}$, and values $\mathbf{V}$ are calculated as:

$$\mathbf{Q} = \mathbf{z}\mathbf{W}_q, \quad \mathbf{K} = \mathbf{e}_{text}\mathbf{W}_k, \quad \mathbf{V} = \mathbf{e}_{text}\mathbf{W}_v, \quad (4)$$

where $\mathbf{W}_q$, $\mathbf{W}_k$, $\mathbf{W}_v$ are the projection matrices. To reinforce the coupling between the physical keywords and target objects, we apply a modulation factor α to the queries within the masked regions. We first denote $\hat{\mathbf{M}} \in \{0, 1\}^S$ as the flattened and interpolated mask aligned with the latent sequence length S , where $S = T' \times H' \times W'$. The modulated queries $\tilde{\mathbf{Q}}$ are defined as:

$$\tilde{\mathbf{Q}}_j = \mathbf{Q}_j \cdot [1 + \hat{\mathbf{M}}_j \cdot (\alpha - 1)], \quad (5)$$

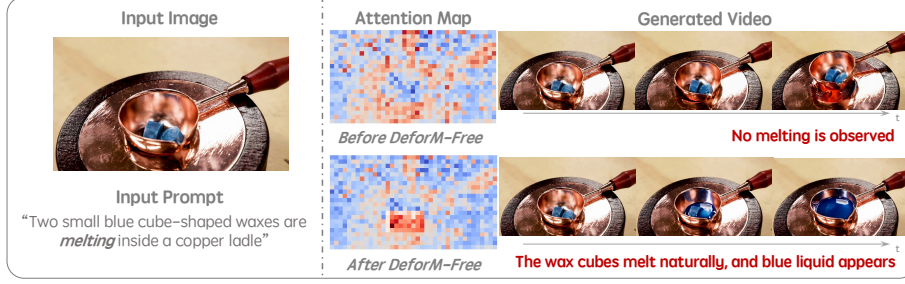


Fig. 3: Motivation and Analysis of DeformM-Free. (Top) Base models exhibit a semantic-spatial gap, where cross-attention for the keyword *melting* is spatially diffused, leading to physical inconsistency. (Bottom) By anchoring attention via DeformM-Free, the attention map becomes more active on the target region, resulting in physically plausible deformation dynamics.

where j is the token index and $\alpha > 1$ is the focus scale. By scaling $\mathbf{Q}$ in the masked area, we effectively amplify the attention scores $\text{Softmax}(\frac{\tilde{\mathbf{Q}}\mathbf{K}^T}{\sqrt{d}})\mathbf{V}$, forcing the model to prioritize the target regions.

As shown in Fig. 3, after applying DeformM-Free, the attention map for *melting* becomes more concentrated on the masked region. The generated video exhibits accurate deformation dynamics, where the wax melts correctly within the boundaries. This validation demonstrates that guiding the model toward physics-critical regions is essential for capturing deformation dynamics. By providing the spatial bias without altering pre-trained weights, DeformM-Free establishes the empirical basis for our subsequent training-based method, DeformM-Injection.

3.4 DeformM-Injection: Training-based Method

While the DeformM-Free method effectively guides the model’s pre-existing knowledge of physical laws, it is inherently limited by the base model’s pre-existing knowledge of physical laws. If the base model lacks a sufficient internal representation of specific deformation dynamics, simply scaling attention scores may result in artifacts or physically inconsistent dynamics. To overcome these limitations, we propose DeformM-Injection, a training-based framework that explicitly injects spatial-temporal priors into the latent space for complex deformation dynamics generation.

To represent the precise physics-critical regions, we design a Dense Mask Modulator $\mathcal{F}_\phi$, serving as an encoder for the spatial-temporal masks. Specifically, given the binary mask $\mathbf{M} \in \{0, 1\}^{T' \times H' \times W'}$ obtained from DeformM-Reason, we first apply a series of 3D convolutional layers to extract hierarchical features while downsampling the mask to match the dimensions of the video latents. The modulator projects the mask into a high-dimensional feature space, generating three decoupled modulation streams: $\Delta\mathbf{k}_{\text{self}}$, $\Delta\mathbf{v}_{\text{self}}$, and $\Delta\mathbf{q}_{\text{cross}}$. To ensure training stability and preserve the foundation model’s generation capability, we

employ a zero-initialization strategy for the final linear layers of the modulator, allowing the network to gradually learn the required offsets.

The core idea of DeformM-Injection lies in the synchronized modulation of both self-attention and cross-attention blocks. Unlike DeformM-Free, which is restricted to enhancing semantic anchoring via training-free cross-attention scaling, this synchronized strategy enables the model to jointly refine external semantic retrieval and internal spatial-temporal consistency through learned feature offsets. We first extend the semantic anchoring concept from DeformM-Free to the training stage. We inject the learned modulation into the cross-attention queries $\mathbf{Q}_{\text{cross}}$, which represent the latent tokens’ active request for physical information from the text prompt. By augmenting $\mathbf{Q}_{\text{cross}}$, the masked regions are empowered to more effectively retrieve physical knowledge from the text embedding. While cross-attention blocks provide semantic guidance, self-attention blocks govern the internal spatial-temporal consistency of the video. We inject mask features into the self-attention keys $\mathbf{K}_{\text{self}}$ and values $\mathbf{V}_{\text{self}}$ within the identified regions. This allows the model to learn localized physical correlations, facilitating the generation of complex and physically plausible deformation dynamics. The modulated attention can be defined as follows:

$$\begin{aligned}\mathbf{K}'_{\text{self}} &= \mathbf{K}_{\text{self}} + \Delta\mathbf{k}_{\text{self}}, \\ \mathbf{V}'_{\text{self}} &= \mathbf{V}_{\text{self}} + \Delta\mathbf{v}_{\text{self}}, \\ \mathbf{Q}'_{\text{cross}} &= \mathbf{Q}_{\text{cross}} + \Delta\mathbf{q}_{\text{cross}},\end{aligned}\tag{6}$$

where $\Delta\mathbf{k}_{\text{self}}, \Delta\mathbf{v}_{\text{self}}, \Delta\mathbf{q}_{\text{cross}} = \mathcal{F}_{\phi}(\mathbf{M})$.

We optimize both the Dense Mask Modulator and a set of Low-Rank Adaptation (LoRA) [11] weights integrated into the transformer backbone. To further prioritize the fidelity of physical deformations, we define a composite loss function. In addition to the standard flow matching loss $\mathcal{L}_{\text{flow}}$ applied to the entire latent sequence, we introduce a region-sensitive loss $\mathcal{L}_{\text{mask}}$ that assigns higher weights to the reconstruction error within physics-critical regions defined by $\mathbf{M}$. The total training objective is:

$$\mathcal{L}_{\text{total}} = (1 - \lambda_{\text{mask}})\mathcal{L}_{\text{flow}} + \lambda_{\text{mask}}\mathcal{L}_{\text{mask}},\tag{7}$$

where λ_{mask} is a balancing hyperparameter. By focusing the optimization on the physics-critical regions, DeformM-Injection enables the model to bridge the gap between abstract physical concepts and concrete pixel-level dynamics, producing videos that are both visually natural and physically plausible.

4 Experiments

This section presents the experimental validation of our proposed methods across multiple dimensions. Specifically, Sec. 4.1 provides the implementation details for both DeformM-Free and DeformM-Injection. Sec. 4.2 introduces the datasets, baselines and the design of evaluation metrics. To demonstrate the effectiveness of our methods, Sec. 4.3 reports both qualitative and quantitative results

compared with SOTA models. The rationality and performance of our designed modules are further analyzed in Sec. 4.4.

4.1 Implementations

We adopt Wan2.2-TI2V-5B [23] as the base image-to-video generation model. To achieve a balance between inference speed and performance, we employ Qwen2.5-VL-7B [29] as the physical reasoning agent and LLMDet [7] as the open-vocabulary object detector for Deform-Reason. Deform-Free is implemented as a training-free strategy by modulating the cross-attention blocks during the inference stage. We set the focus scale α to 1.4 to provide precise spatial-temporal guidance while maintaining generation quality.

For Deform-Injection, we fine-tune Wan2.2-TI2V-5B using LoRA with a rank of 256. The training process utilizes a batch size of 64 with a learning rate of 1×10^{-6} for the LoRA weights and a higher rate of 2×10^{-6} for the modulator. The balancing hyperparameter λ_{mask} is set to 0.3 to prioritize the reconstruction of physics-critical regions.

4.2 Evaluation Settings

Datasets. We conduct our experiments using a specialized subset of the WISA-80K [24] dataset. To ensure training quality, we select approximately 6k high-quality deformation dynamics samples, including melting, squeezing, fracturing, and other scenarios. Each sample consists of a ground-truth video captured in real-world environments with a corresponding text caption describing the physical action. For evaluation, we utilize a separate test set comprising 320 samples across diverse scenarios for visual quality and randomly select 160 samples for physical consistency evaluation. These evaluation samples are excluded from the training set across diverse deformation categories.

Baselines. To evaluate the performance of our framework, we compare it against several open-source SOTA models that have demonstrated robust image-to-video generation capabilities. These competitive baselines include Hunyuan-I2V [14], CogVideoX1.5-I2V [31], Wan2.2-14B-I2V [23] and MAGI-1 [21]. We selected these models for their outstanding video generation performance and widespread usage.

Evaluation Metrics. For a comprehensive evaluation, we assess the generated results across both visual quality and physical consistency dimensions. Regarding visual quality, we adopt VBench [12] as the foundational benchmark following previous works [34, 35]. Considering the unique nature of deformation dynamics where foreground objects undergo significant structural changes and may affect the background, we exclude *subject consistency* and *background consistency* as these attributes are inherently compromised during such processes. Instead, we retain six other metrics that comprehensively reflect video quality: *motion smoothness*, *temporal flickering*, *dynamic degree*, *aesthetic quality*, *imaging quality* and the aggregate overall *quality score*.

Table 1: Quantitative comparison of DeformM-Injection with state-of-the-art image-to-video generation models. Blue and yellow shaded columns represent the aggregate scores for visual quality and physical consistency, respectively. The best results are highlighted in bold.

Method	Visual Quality						Physical Consistency			
	Temporal Flickering ↑	Motion Smoothness ↑	Dynamic Degree ↑	Aesthetic Quality ↑	Imaging Quality ↑	Quality Score ↑	PC ↑	SA ↑	LDF ↑	Physical Score ↑
CogVideoX1.5-I2V	98.25	98.84	48.32	45.68	65.92	71.40	4.07	3.62	3.24	3.64
Hunyuan-I2V	99.30	99.54	27.74	46.15	64.80	67.50	3.86	3.34	3.01	3.40
Wan2.2-I2V-14B	97.42	98.34	46.07	46.61	67.47	71.18	4.28	3.91	3.65	3.95
MAGI-1	99.09	99.41	28.03	45.97	63.79	67.26	3.93	3.39	3.01	3.44
DeformM-Injection (Ours)	97.24	98.40	54.74	46.61	68.84	73.17	4.59	4.46	4.16	4.40

For physical consistency, we utilize Qwen3-VL-32B [28] as a powerful evaluator, considering its exceptional multimodal physical evaluation capabilities. The generated videos and corresponding captions are provided as inputs to score the results across three distinct dimensions using a scale from 1 to 5. The first dimension is *Physical Commonsense* (PC), which evaluates the adherence to the general physics of the entire video. The second dimension is *Semantic Adherence* (SA), assessing how well the video follows the textual instructions. The third dimension is *Local Deformation Fidelity* (LDF), which focuses on the accuracy and realism of local deformation dynamics. For each dimension, the score ranges from 1 representing *Poor* and 5 representing *Excellent*, providing a detailed quantitative measure of physical realism. We define the *Physical Score* as the arithmetic mean of these three dimensions, providing a comprehensive assessment of the overall physical plausibility.

4.3 Comparisons

Quantitative Results. Summarized in Tab. 1, the quantitative comparison results demonstrate the superior performance of our proposed method in both visual quality and physical consistency. For visual quality, our method achieves the highest overall quality score and demonstrates a significant advantage in the dynamic degree. Although some baselines, like Hunyuan-I2V, show high performance in temporal flickering and motion smoothness, they tend to generate nearly static scenes with minimal movement. In contrast, our approach generates obvious and realistic deformation dynamics with competitive temporal consistency and motion smoothness. Regarding physical consistency, our method shows superior results across all dimensions with the most notable advantages in LDF. These results demonstrate the effectiveness of the integration of VLM-guided physical reasoning and spatial-temporal mask injection.

Qualitative Results. In this section, we present qualitative results to demonstrate the generation quality and physical plausibility of our method. Fig. 4 presents diverse deformation scenarios reflecting the physical consistency and

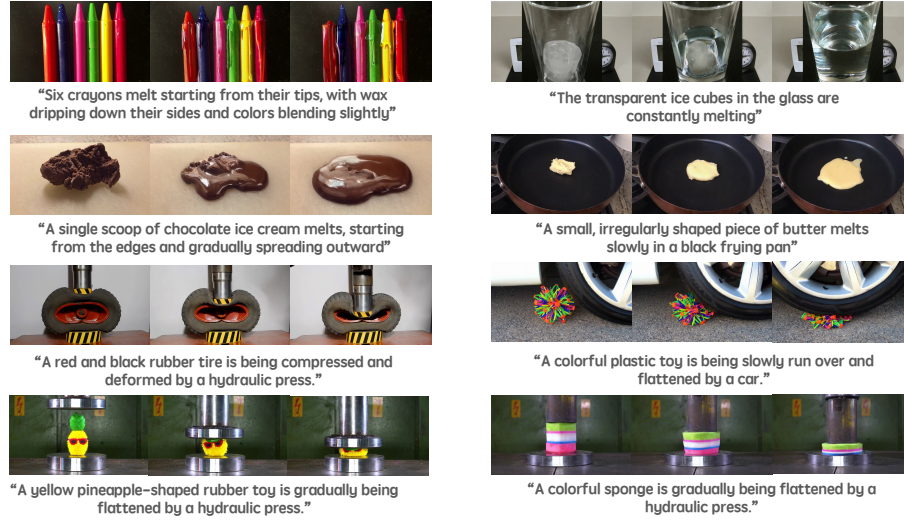


Fig. 4: Visualization results. Visualization of diverse deformation scenarios generated by Deform-Injection.

dynamics achieved by our work. Furthermore, Fig. 5 provides a visual comparison between Deform and baselines across two representative deformation types involving melting and squeezing. In the melting examples, baseline models suffer from various limitations such as poor object understanding, failure to generate the melting effect, or extraneous artifacts. Regarding the squeezing scenarios, the baseline models exhibit significant issues including disappearing objects, a lack of squeezing dynamics, and inconsistent object colors. In contrast, our method demonstrates superior physical plausibility and vivid local deformation dynamics, indicating the effectiveness of our framework.

4.4 Ablation Studies

To evaluate the effectiveness of each component, we conduct an ablation study against the vanilla Wan2.2-I2V-5B and standard supervised fine-tuning (SFT). As shown in Tab. 2, the vanilla model and SFT method struggle to generate accurate deformation dynamics with lower dynamic degree and physical scores. To assess the impact of Deform-Reason, we replace our VLM-guided masks with randomly generated spatial-temporal masks. The performance degradation confirms that precise localization of physics-critical regions is essential for effective guidance. Notably, Deform-Free achieves strong performance in physical plausibility without any additional training, even surpassing the SFT method.

Furthermore, we independently modulate cross-attention and self-attention, denoted as Deform-Cross and Deform-Self, respectively, as shown in Tab. 3. We observe that cross-attention modulation, Deform-Cross, benefits semantic adherence and physical commonsense by aligning deformation information with

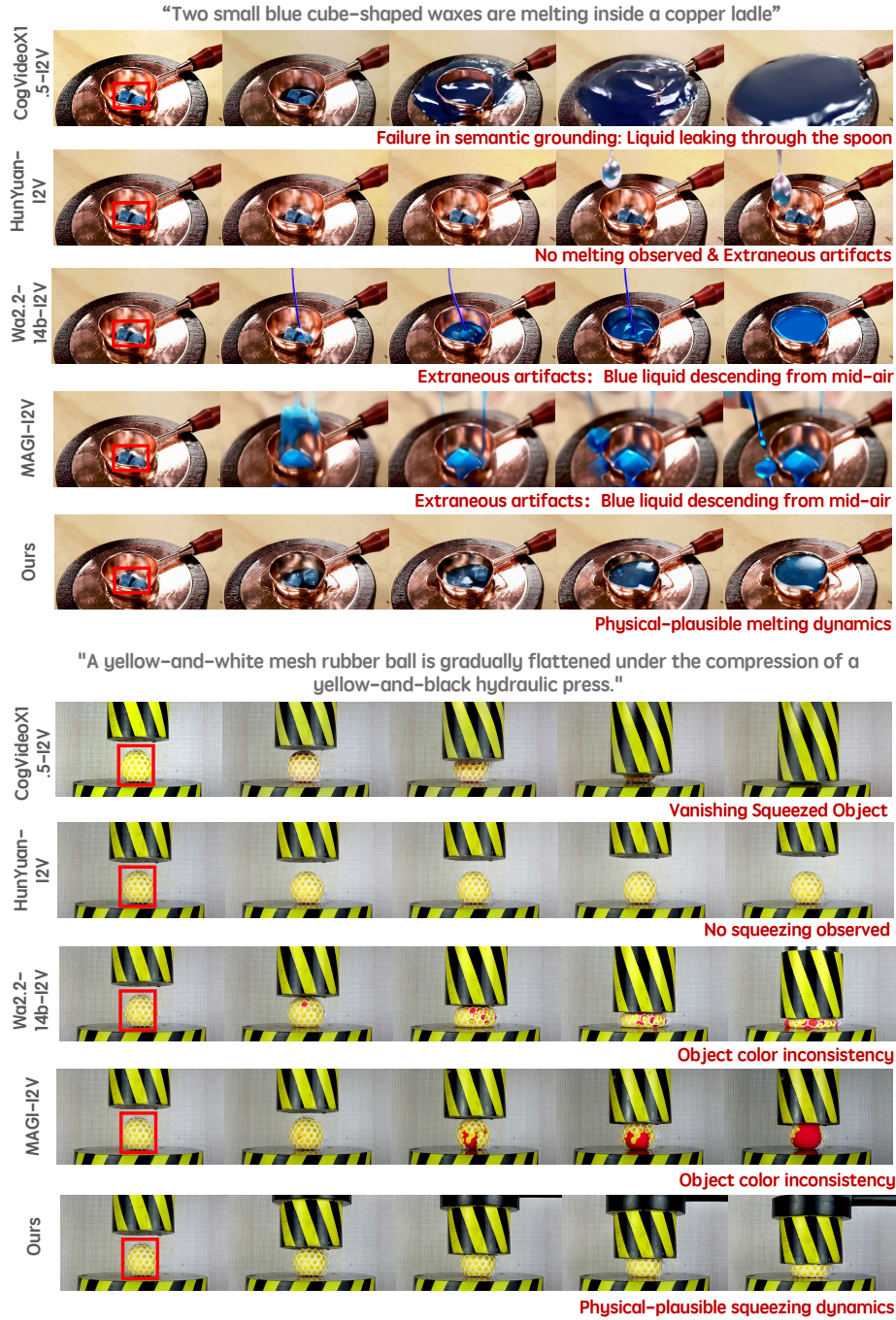


Fig. 5: Qualitative comparisons. DeformM presents superior physical plausibility and vivid local deformation dynamics. The red bounding boxes indicate the physics-critical regions.

Table 2: Ablation study of different configurations and baselines. Blue and yellow shaded columns represent the aggregate scores for visual quality and physical consistency, respectively. The best results are highlighted in bold.

Method	Visual Quality						Physical Consistency			
	Temporal Flickering ↑	Motion Smoothness ↑	Dynamic Degree ↑	Aesthetic Quality ↑	Imaging Quality ↑	Quality Score ↑	PC ↑	SA ↑	LDF ↑	Physical Score ↑
Vanilla Wan2.2	98.12	98.85	42.61	46.28	67.40	70.65	3.97	3.31	3.19	3.49
SFT	97.71	98.54	44.77	45.99	67.65	70.93	4.19	3.69	3.41	3.76
W/O Deform-Reason	97.70	98.62	47.58	45.60	67.52	71.40	4.22	3.79	3.45	3.82
DeformM-Free	97.21	98.27	51.57	46.47	66.41	71.99	4.28	3.86	3.53	3.89
DeformM-Injection	97.24	98.40	54.74	46.61	68.84	73.17	4.59	4.46	4.16	4.40

Table 3: Ablation study of different attention injection mechanisms. Blue and yellow shaded columns represent the aggregate scores for visual quality and physical consistency, respectively. The best results are highlighted in bold.

Method	Visual Quality						Physical Consistency			
	Temporal Flickering ↑	Motion Smoothness ↑	Dynamic Degree ↑	Aesthetic Quality ↑	Imaging Quality ↑	Quality Score ↑	PC ↑	SA ↑	LDF ↑	Physical Score ↑
DeformM-Cross	97.37	98.43	53.23	46.42	67.74	72.64	4.42	4.01	3.83	4.09
DeformM-Self	97.68	98.62	53.16	46.50	68.21	72.83	4.22	3.73	3.33	3.76
DeformM-Injection	97.24	98.40	54.74	46.61	68.84	73.17	4.59	4.46	4.16	4.40

spatial-temporal regions. Injection into self-attention, DeformM-Self, contributes more to temporal smoothness and imaging quality. DeformM-Injection achieves the highest performance, especially in LDF and overall physical score.

5 Conclusion

In this paper, we propose DeformM, a reasoning-guided image-to-video framework for physics-aware deformation synthesis. To enable physical reasoning in deformation generation, we introduce a VLM-guided module to analyze object physical properties in multimodal inputs and ground physics-critical regions as spatial-temporal masks. Furthermore, we develop training-free and training-based mask injection strategies to steer the model’s generation focus. Extensive experiments demonstrate the effectiveness of our framework, especially in local deformation fidelity and dynamic degree. Our work highlights the importance of incorporating physical reasoning and spatial-temporal physical priors into generative frameworks, providing a promising pathway for physics-aware video generation.

Limitations and Future Work. Despite promising results, DeformM remains limited by the localization accuracy of the DeformM-Reason module, which may produce suboptimal spatial-temporal masks in complex scenes. Moreover, as DeformM builds on LoRA fine-tuning of a foundation model, its generation quality and scene diversity are partly constrained by the base model. These limitations may be alleviated by future advances in grounding, reasoning, and foundation video generation.

Acknowledgements

This work is supported by Shanghai Artificial Intelligence Laboratory. We gratefully acknowledge the support and research resources provided by Shanghai Artificial Intelligence Laboratory throughout this project.

References

1. Albergo, M.S., Vanden-Eijnden, E.: Building normalizing flows with stochastic interpolants. In: The Eleventh International Conference on Learning Representations (2023)
2. Bai, J., Bai, S., Chu, Y., Cui, Z., Dang, K., Deng, X., Fan, Y., Ge, W., Han, Y., Huang, F., Hui, B., Ji, L., Li, M., Lin, J., Lin, R., Liu, D., Liu, G., Lu, C., Lu, K., Ma, J., Men, R., Ren, X., Ren, X., Tan, C., Tan, S., Tu, J., Wang, P., Wang, S., Wang, W., Wu, S., Xu, B., Xu, J., Yang, A., Yang, H., Yang, J., Yang, S., Yao, Y., Yu, B., Yuan, H., Yuan, Z., Zhang, J., Zhang, X., Zhang, Y., Zhang, Z., Zhou, C., Zhou, J., Zhou, X., Zhu, T.: Qwen technical report. arXiv preprint arXiv:2309.16609 (2023)
3. Chen, H.H., Huang, H., Chen, Q., Yang, H., Lim, S.N.: Hierarchical fine-grained preference optimization for physically plausible video generation. In: The Thirty-ninth Annual Conference on Neural Information Processing Systems (2025)
4. Chen, R.T., Rubanova, Y., Bettencourt, J., Duvenaud, D.K.: Neural ordinary differential equations. *Advances in neural information processing systems* **31** (2018)
5. Cong, X., Yang, H., Wang, A., Wang, Y., Yang, Y., Zhang, C., Ma, C.: Viva: Vlm-guided instruction-based video editing with reward optimization. arXiv preprint arXiv:2512.16906 (2025)
6. Davtyan, A., Sameni, S., Favaro, P.: Efficient video prediction via sparsely conditioned flow matching. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 23263–23274 (2023)
7. Fu, S., Yang, Q., Mo, Q., Yan, J., Wei, X., Meng, J., Xie, X., Zheng, W.S.: Llm-det: Learning strong open-vocabulary object detectors under the supervision of large language models. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 14987–14997 (2025)
8. Gillman, N., Herrmann, C., Freeman, M., Aggarwal, D., Luo, E., Sun, D., Sun, C.: Force prompting: Video generation models can learn and generalize physics-based control signals. In: The Thirty-ninth Annual Conference on Neural Information Processing Systems (2025)
9. He, H., Zhang, Y., Lin, L., Xu, Z., Pan, L.: Pre-trained video generative models as world simulators. In: ICLR 2025 Workshop on World Models: Understanding, Modelling and Scaling (2025)
10. Ho, J., Salimans, T.: Classifier-free diffusion guidance. arXiv preprint arXiv:2207.12598 (2022)
11. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: LoRA: Low-rank adaptation of large language models. In: International Conference on Learning Representations (2022)
12. 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)

13. Jin, Y., Sun, Z., Li, N., Xu, K., Xu, K., Jiang, H., Zhuang, N., Huang, Q., Song, Y., MU, Y., Lin, Z.: Pyramidal flow matching for efficient video generative modeling. In: The Thirteenth International Conference on Learning Representations (2025)
14. 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)
15. Li, C., Michel, O., Pan, X., Liu, S., Roberts, M., Xie, S.: PISA experiments: Exploring physics post-training for video diffusion models by watching stuff drop. In: Forty-second International Conference on Machine Learning (2025)
16. Li, Z., Yu, H.X., Liu, W., Yang, Y., Herrmann, C., Wetzstein, G., Wu, J.: Wonderplay: Dynamic 3d scene generation from a single image and actions. In: Proceedings of the IEEE/CVF international conference on computer vision (2025)
17. Lipman, Y., Chen, R.T.Q., Ben-Hamu, H., Nickel, M., Le, M.: Flow matching for generative modeling. In: The Eleventh International Conference on Learning Representations (2023)
18. Liu, S., Ren, Z., Gupta, S., Wang, S.: Physgen: Rigid-body physics-grounded image-to-video generation. In: European Conference on Computer Vision ECCV (2024)
19. 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)
20. Tan, X., Jiang, Y., Li, X., Xie, T., Zong, Z., Yang, Y., Jiang, C.: Physmotion: Physics-grounded dynamics from a single image. In: Thirteenth International Conference on 3D Vision (2026)
21. Teng, H., Jia, H., Sun, L., Li, L., Li, M., Tang, M., Han, S., Zhang, T., Zhang, W., Luo, W., et al.: Magi-1: Autoregressive video generation at scale. arXiv preprint arXiv:2505.13211 (2025)
22. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. *Advances in neural information processing systems* **30** (2017)
23. Wan, T., Wang, A., Ai, B., Wen, B., Mao, C., Xie, C.W., Chen, D., Yu, F., Zhao, H., Yang, J., Zeng, J., Wang, J., Zhang, J., Zhou, J., Wang, J., Chen, J., Zhu, K., Zhao, K., Yan, K., Huang, L., Feng, M., Zhang, N., Li, P., Wu, P., Chu, R., Feng, R., Zhang, S., Sun, S., Fang, T., Wang, T., Gui, T., Weng, T., Shen, T., Lin, W., Wang, W., Wang, W., Zhou, W., Wang, W., Shen, W., Yu, W., Shi, X., Huang, X., Xu, X., Kou, Y., Lv, Y., Li, Y., Liu, Y., Wang, Y., Zhang, Y., Huang, Y., Li, Y., Wu, Y., Liu, Y., Pan, Y., Zheng, Y., Hong, Y., Shi, Y., Feng, Y., Jiang, Z., Han, Z., Wu, Z.F., Liu, Z.: Wan: Open and advanced large-scale video generative models. arXiv preprint arXiv:2503.20314 (2025)
24. Wang, J., Ma, A., Cao, K., Zheng, J., Feng, J., Zhang, Z., Pang, W., Liang, X.: WISA: World simulator assistant for physics-aware text-to-video generation. In: The Thirty-ninth Annual Conference on Neural Information Processing Systems (2025)
25. Wang, X., Zhu, Z., Huang, G., Wang, B., Chen, X., Lu, J.: Worlddreamer: Towards general world models for video generation via predicting masked tokens. arXiv preprint arXiv:2401.09985 (2024)
26. 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 diffusion models. *International Journal of Computer Vision* (2024)

27. Wu, J., Tang, C., Wang, J., Zeng, Y., Li, X., Tong, Y.: Diffsensei: Bridging multi-modal llms and diffusion models for customized manga generation. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 28684–28693 (2025)
28. Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., Zheng, C., Liu, D., Zhou, F., Huang, F., Hu, F., Ge, H., Wei, H., Lin, H., Tang, J., Yang, J., Tu, J., Zhang, J., Yang, J., Yang, J., Zhou, J., Zhou, J., Lin, J., Dang, K., Bao, K., Yang, K., Yu, L., Deng, L., Li, M., Xue, M., Li, M., Zhang, P., Wang, P., Zhu, Q., Men, R., Gao, R., Liu, S., Luo, S., Li, T., Tang, T., Yin, W., Ren, X., Wang, X., Zhang, X., Ren, X., Fan, Y., Su, Y., Zhang, Y., Zhang, Y., Wan, Y., Liu, Y., Wang, Z., Cui, Z., Zhang, Z., Zhou, Z., Qiu, Z.: Qwen3 technical report. arXiv preprint arXiv:2505.09388 (2025)
29. Yang, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Li, C., Liu, D., Huang, F., Wei, H., Lin, H., Yang, J., Tu, J., Zhang, J., Yang, J., Yang, J., Zhou, J., Lin, J., Dang, K., Lu, K., Bao, K., Yang, K., Yu, L., Li, M., Xue, M., Zhang, P., Zhu, Q., Men, R., Lin, R., Li, T., Xia, T., Ren, X., Ren, X., Fan, Y., Su, Y., Zhang, Y., Wan, Y., Liu, Y., Cui, Z., Zhang, Z., Qiu, Z.: Qwen2.5 technical report. arXiv preprint arXiv:2412.15115 (2024)
30. Yang, X., Li, B., Zhang, Y., Yin, Z., Bai, L., Ma, L., Wang, Z., Cai, J., Wong, T.T., Lu, H., Jia, X.: Vlpp: Towards physically plausible video generation with vision and language informed physical prior. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 12360–12370 (October 2025)
31. Yang, Z., Teng, J., Zheng, W., Ding, M., Huang, S., Xu, J., Yang, Y., Hong, W., Zhang, X., Feng, G., Yin, D., Yuxuan.Zhang, Wang, W., Cheng, Y., Xu, B., Gu, X., Dong, Y., Tang, J.: Cogvideox: Text-to-video diffusion models with an expert transformer. In: The Thirteenth International Conference on Learning Representations (2025)
32. Ye, W., Liu, Z., Yuwei, G., Yuan, T., Su, Y., Fang, B., Zhao, C., Liu, Q., Wang, L.: GenPilot: A multi-agent system for test-time prompt optimization in image generation. In: Christodoulopoulos, C., Chakraborty, T., Rose, C., Peng, V. (eds.) Findings of the Association for Computational Linguistics: EMNLP 2025. pp. 929–958. Association for Computational Linguistics, Suzhou, China (Nov 2025)
33. Zhang, D., He, L., Yan, R., Shen, F., Tang, J.: R-genie: Reasoning-guided generative image editing (2025)
34. Zhang, Q., Gong, B., Tan, S., Zhang, Z., Shen, Y., Zhu, X., Li, Y., Yao, K., Shen, C., Zou, C.: Physrv: Physics-aware unified reinforcement learning for video generative models. arXiv preprint arXiv:2601.11087 (2025)
35. Zhao, Y., Li, H., He, X., Wu, B.: Phyrpr: Training-free physics-constrained video generation (2026)