POLITECNICO DI TORINO
Repository ISTITUZIONALE

MOTIONCRAFT: Physics-based Zero-Shot Video Generation

Original
MOTIONCRAFT: Physics-based Zero-Shot Video Generation / Montanaro, A., SAVANT AIRA, L., Aiello, E., Valsesia, D., Magli, E.. - 37:(2024). (38th Conference on Neural Information Processing Systems, NeurIPS 2024 Vancouver (Can) 10 - 15 December 2024).

Availability:
This version is available at: 11583/3000132 since: 2025-05-14T14:01:21Z

Publisher:
Neural information processing systems foundation

Published
DOI:

Terms of use:

This article is made available under terms and conditions as specified in the corresponding bibliographic description in the repository

Publisher copyright

(Article begins on next page)

09 October 2026

MOTIONCRAFT: Physics-based Zero-Shot Video Generation

Antonio Montanaro* **Luca Savant Aira***

Emanuele Aiello, **Diego Valsesia** **Enrico Magli**

Politecnico di Torino
{name.surname}@polito.it

Abstract

Generating videos with realistic and physically plausible motion is one of the main recent challenges in computer vision. While diffusion models are achieving compelling results in image generation, video diffusion models are limited by heavy training and huge models, resulting in videos that are still biased to the training dataset. In this work we propose MotionCraft, a new zero-shot video generator to craft physics-based and realistic videos. MotionCraft is able to warp the noise latent space of an image diffusion model, such as Stable Diffusion, by applying an optical flow derived from a physics simulation. We show that warping the noise latent space results in coherent application of the desired motion while allowing the model to generate missing elements consistent with the scene evolution, which would otherwise result in artefacts or missing content if the flow was applied in the pixel space. We compare our method with the state-of-the-art Text2Video-Zero reporting qualitative and quantitative improvements, demonstrating the effectiveness of our approach to generate videos with finely-prescribed complex motion dynamics.

1 Introduction

As human beings, we have always exploited our creativity to generate art, in different forms such as visual art, music or poetry. In vision, we are often inspired by the natural world since our visual system continuously acquire images perceived as a video sequence. Indeed, videos or movies are one of the best visual stimuli since they contain images, motion and audio.

Recent generative models for still images based on diffusion models [25, 28, 29] achieved remarkable results with quality almost indistinguishable from real images. It is therefore clear that the next big goal is video generation. However, it seems that including the dimension of time remains challenging. Some works such as Sora [5] achieve astonishing temporal consistency and photorealism at the expense of enormous computational and data requirements. Moreover, we argue that fine-grained control over the motion dynamics is impossible with a simple text prompt. If one wants to synthesize a video according to some precise physical dynamics, they would not be able to do it with current models. Interestingly, explicitly controlling the motion dynamics also allows to decouple temporal evolution from content generation. Indeed, explicitly injecting the physics of the

*indicates equal contribution.
Project page: https://mezzelfo.github.io/MotionCraft/.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

Figure 1: Melting man simulation. Top: MOTIONCRAFT; Bottom: T2V0 [20]. MOTIONCRAFT uses a fluid dynamics simulation to warp noise latents and synthetize video frames. T2V0 is unable to simulate the evolution of the melting statue and simply moves the object towards the bottom of the frame.

real world as motion dynamics allows to develop more parsimonious models, that do not need to brute-force learn them from data.

For this reason, in this paper, we investigate the possibility to create a zero-shot video generation model that only requires a pretrained still image generator and knowledge of physical laws regarding motion. Indeed, since videos are temporal sequences of images correlated by physical laws, we only need to devise a way to include physical laws in the diffusion prior to animate a starting image. We thus advocate for physics simulators as appropriate sources of motion, output as a sequence of optical flows, while also being completely user-controllable, plausible, and explainable.

We propose MOTIONCRAFT, a physics-based zero-shot video generator that uses optical flow extracted from a physical simulation to warp the noise latent space of a pretrained image diffusion model to generate videos with complex dynamics without the need to train anything. While using a projection of motion onto the camera plane as a pixelwise displacement field (optical flow) may seem limiting due to the fact that, if applied in the pixel space, it would not be able to synthesise novel coherent content but only displace pixels, the trick lies in its application in the noise latent domain. Backed by evidence that motion vectors correlate between pixel and noise space, warping of the latter by means that MOTIONCRAFT allows to simultaneously apply the desired motion and exploit the powerful image prior of the generative model. This is capable of adapting the scene to the prescribed motion without significant artefacts, generate novel content and shows impressive global consistency (reflections, illumination, etc., consistent with the desired evolution).

We present quantitative and qualitative experimental results where we show that our zero-shot MOTIONCRAFT is capable of synthesising realistic videos with finely controlled temporal evolution governed by fluid-dynamics equations, rigid body physics, and multi-agent interaction models, while zero-shot state-of-art techniques cannot.

2 Related work

Diffusion Based Video Generation Video Generation [2] is a longstanding problem in computer vision aiming to learn the distribution of and synthesise realistic videos. Recently, text-based Denoising Diffusion Probabilistic Models (DDPM) [28, 31] have been studied to tackle this challenge delivering impressive results. These approaches include Sora [5], Video Diffusion models [17], Imagen-video [16] and Align your Latents [3]. They require sophisticated spatio-temporal denoising architectures at the expense of huge computational requirements and large amounts of paired text-video data for training. To reduce the data requirements, different approaches investigate few-shot and unsupervised learning techniques. Make-a-Video [27] proposes an unsupervised training with only videos, coupled with a retrieval strategy to sample using text. On the other hand, Ni et al. [22] train a diffusion-based optical flow generator that outputs a flow conditioned on a reference image and a textual prompt, that reduces the computational burden of generating videos by training the diffusion process on small flow fields. Differently from them, our approach is zero-shot and we do not train anything.

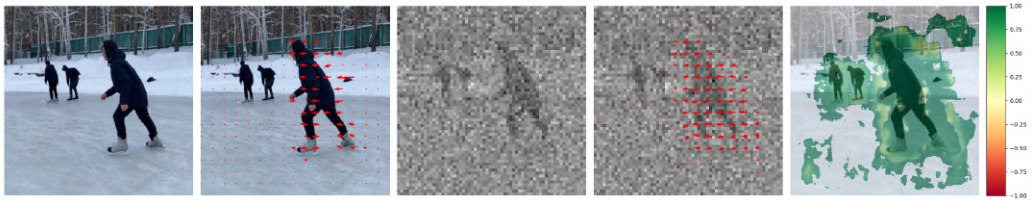

Figure 2: A qualitative example of the image and latent flows correlation. This figure shows, from left to right, (a) the first RGB frame, (b) the second RGB frame superimposed with the estimated flow in the RGB domain, (c) the first latent frame, (d) the second latent frame superimposed with the estimated flow in the latent domain and (e) the correlation map of the two non-zero flows.

To the best of our knowledge, Text-to-video-Zero [20] and Generative Rendering [6] are the only zero-shot video generators. However, Generative Rendering (concurrent work, with no code available) has significant extra requirements beyond Stable Diffusion (SD) as image generator, in the form of a depth-conditioned ControlNet [36], and a 3D mesh manually animated, leveraging UV maps to render the scene. Moreover, Generative Rendering cannot render fluids, since they are difficult to represent as 3D meshes.

In this paper, we compare our method to Text-to-video-Zero (T2V0), as zero-shot video generator baseline. T2V0 applies a constant shift (with a fixed direction) to the initial latent noise of SD, sampling each frame sequentially by means of DDPM. As shown in our work, since the motion in the noise latent space directly translates into the motion of the pixel space, the generated videos result in a overall shift in the same fixed direction. The largest part of the motion is caused by the stochastic fluctuations of the DDPM sampling strategy leading to unnatural motion and inconsistency of the objects in the different frames. On the contrary, in this work, we avoid the use of a constant warping operation derived from physics simulation flows in the latent space in order to incorporate complex motion dynamics.

Diffusion Based Video and Image Editing Recently, different methods exploit the prior of text-to-image diffusion models for video editing. In particular, Tune-A-Video [35] finetunes a text-to-image diffusion model to edit a video. They start from the inverted frames in the latent space and use the text prompt as an editing tool. Pix2Video [7] employs a self-attention injection mechanism to edit videos using a pretrained image diffusion model.

Other methods use the optical flow to edit reference images or videos. Motion Guidance [10] leverages a user defined optical flow that allow zero-shot image editing. It works by guiding the diffusion sampling process with the gradient from a pretrained optical flow network via a guidance loss. LatentWarp [1] and TokenFlow [11], use an optical flow estimated from a reference video to warp the latent space of the diffusion model to achieve consistent editing. These methods leverage both diffusion models priors and other components such as ControlNet for structural control, and trained flow estimators such as RAFT [32]. Alternatively, we propose a zero-shot video generation method, using only vanilla SD. This means that MOTIONCRAFT does not require a reference video but it can animate an image, generated by the SD model or obtained by inverting a real one. Moreover, the physics simulations allow to generate different videos from the same starting image.

3 Method

This section describes MOTIONCRAFT, a zero-shot video generation method, where the meaning of "zero-shot" is twofold: we do not train or finetune any component of the text-to-image diffusion model, nor we do not use reference video or optical flow estimators as starting point. In the following, we used used Stable Diffusion as pretrained text-to-image model.

3.1 Optical Flow is preserved in the Latent Space of Stable Diffusion

Our proposed method stems from a key observation: the optical flow estimated between two frames in the pixel space is correlated with the flow estimated between the corresponding noise latent representations of SD. We conjecture that this is related to the specific design of the SD variational

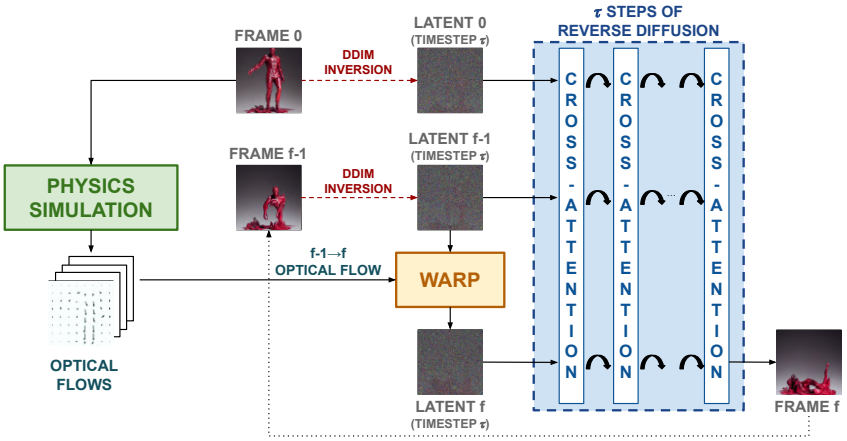

Figure 3: MOTIONCRAFT overview. A video is generated from a starting image using a pretrained still image generative model by warping noise latents according to an optical flow description of the motion to be synthesised.

auto-encoder and denoiser architectures. In fact, by largely using convolution operations, they enforce a locality prior which preserves spatial information to some extent.

In order to empirically investigate this phenomenon, we conducted a quantitative experiment using the MSU Video Frame Interpolation Benchmark dataset [12], considering only real videos. For each pair of consecutive video frames, the following steps have been taken. We first estimate the optical flow in the RGB space by using a well-established method, based on the Gunnar Farneback's algorithm, provided by OpenCV [19]. Then, we compute the noise latent representations of the two frames, first encoding the image in the variational autoencoder (VAE) of SD at timestep $\tau = 0$, followed by DDIM inversion [29] up to timestep $\tau = 400$ (same value for all experiments in this work, empirically determined). Finally, a correlation coefficient based on cosine similarity is computed between the optical flows estimated in the RGB and noise latent spaces. The resulting correlations are then averaged across all pairs of consecutive frames in the dataset, obtaining an average value of 0.727, which indicates a strong correlation between the optical field in the RGB and noise latent domains. An example of this experiment is presented in Fig. 2, showcasing the two estimated flows in the image and latent space and their correlation.

3.2 Physics-based zero shot video generation

Based on the analysis presented in the previous section, we propose a novel zero-shot video generation method, named MOTIONCRAFT, where an image (real or generated), serving as a starting frame I^0, is animated according to a physical simulation, by means of a (possibly time-varying) optical-flow generator $\mathcal{W}$ in the noise latent space. The outcome is a video made of N frames $I^0, \ldots, I^{N-1}$ that follows the motion prescribed by the physical simulation and evolves the content of the first frame coherently. Inspired by the previous observation, this animation is obtained by warping the noisy latent representation of an image in the latent diffusion space. Regarding the physics simulation for the optical flow generation, we use different libraries to simulate different physics, as explained in the experimental section, such as fluid dynamics, rigid motion and multi-agent systems. It is also possible, albeit not shown in this paper, to use animation software to generate the required optical flows.

Fig. 3 illustrates an overview of MOTIONCRAFT highlighting the autoregressive generation of the video. At each iteration $f \geq 1$, the frame I^f is generated using only the information contained in the first frame I^0 and the previous frame I^{f-1}. Given this Markovian structure, MOTIONCRAFT is characterized by $\mathcal{O}(1)$ space complexity and $\mathcal{O}(N)$ time complexity with respect to the total number N of frames to be generated. More in detail, first, the two RGB frames I^0, I^{f-1} are encoded into the latent space and they are independently inverted with the reversed DDIM sampling scheme up to a fixed diffusion timestep τ, obtaining z_τ^0, and z_τ^{f-1}, respectively. Then, the optical flow warping operator $\mathcal{W}^{f-1 \to f}$ prescribed by the physical simulation is applied to z_τ^{f-1}, obtaining ζ_τ^f.

Algorithm 1: Pseudocode of MOTIONCRAFT

Input : $I^0, \mathcal{W}, \eta, \mathcal{P}, \mathcal{P}_\emptyset$
Output : $I^0, \dots, I^{N-1}$

1 **for** $f = 1$ **to** $N-1$ **do**
2 $\quad z_0^{f-1} = \mathcal{E}(I^{f-1})$; // Encode the frame
3 $\quad$ **for** $t = 0$ **to** $\tau - 1$ **do** // Inversion loop
4 $\quad\quad \hat{\epsilon} \leftarrow \epsilon_t(z_t^{f-1}, \mathcal{P}; \{z_t^{f-1}\})$; // Self-Attention, No MCFA
5 $\quad\quad z_{t+1}^{f-1} \leftarrow \text{DDIMInversion}_{t \to t+1}(z_t^{f-1}, \hat{\epsilon}, 0)$; // $\eta = 0 \iff$ DDIM
6 $\quad$ **end**
7 $\quad \zeta_\tau^f = \mathcal{W}^{f-1 \to f}(z_\tau^{f-1})$; // Warp the latent
8 $\quad$ **for** $t = \tau - 1$ **to** 0 **do** // Generation loop
9 $\quad\quad \hat{\epsilon}_\mathcal{P} \leftarrow \epsilon_t(\zeta_{t+1}^f, \mathcal{P}; \{z_t^0, z_t^{f-1}\})$; // MCFA with I^0 and I^{f-1}
10 $\quad\quad \hat{\epsilon}_\emptyset \leftarrow \epsilon_t(\zeta_{t+1}^f, \mathcal{P}_\emptyset; \{z_t^0, z_t^{f-1}\})$; // MCFA with I^0 and I^{f-1}
11 $\quad\quad \hat{\epsilon} \leftarrow \hat{\epsilon}_\emptyset + \gamma(\hat{\epsilon}_\mathcal{P} - \hat{\epsilon}_\emptyset)$; // Classifier-free guidance
12 $\quad\quad \zeta_t^f \leftarrow \text{DDIM}_{t+1 \to t}(\zeta_{t+1}^f, \hat{\epsilon}, \eta^f)$; // Perform Spatial-η
13 $\quad$ **end**
14 $\quad I^f \leftarrow \mathcal{D}(\zeta_0^f)$; // Decode the latent
15 **end**
16 **return** $I^0, \dots, I^{N-1}$

Finally, the next RGB frame I^f is generated by performing τ steps of reverse diffusion using the DDIM sampling scheme with a novel cross-frame attention mechanism and a novel spatial noise map η^f weighting technique, explained below. Furthermore, we exploit the classifier-free guidance (CFG) technique for generation proposed in [14], with $\mathcal{P}$ and $\mathcal{P}_\emptyset$ being the positive and negative prompt, respectively, and $\gamma > 1$ being the strength of the CFG. More details can be found in the Appendix A.

Algorithm 1 reports the pseudocode of MOTIONCRAFT. Lines $2-6$ include the DDIM inversion up to timestep τ. Starting current frame I^{f-1} that was previously generated, in line 2 we embed it with the VAE encoder $\mathcal{E}$, obtaining z_0^{f-1}. Then we apply DDIM inversion on z_0^{f-1} for τ timesteps (line $3-6$). This involves the UNet with the standard self-attention (note the repetition of the noisy latent z_t^{f-1}) and the positive prompt $\mathcal{P}$. As briefly reported in [21], we have also experienced that DDIM inversion is not compatible with CFG; hence, during the inversion, we do not use the negative prompt $\mathcal{P}_\emptyset$. The resulting estimated noise is used in line 5 for applying the DDIM inversion step (note that the $\eta = 0$, so pure DDIM is performed). Upon completion of the DDIM inversion process, we obtain z_τ^{f-1}, the noisy latent corresponding to the frame I^{f-1}.

In line 7, the optical flow warping operator $\mathcal{W}^{f-1 \to f}$ is applied to the noisy latent of the current frame z_τ^{f-1} to obtain a new noisy latent ζ_τ^f that will generate the successive frame. Finally, in lines $8-14$, the frame is generated. During this generation phase we use CFG to increase the quality of the generated images, hence also the negative prompt $\mathcal{P}_\emptyset$ is used. To create new content while preserving the original image, we propose two direct generalization of two known techniques: the multiple cross-frame attention (MCFA) mechanism and a spatial noise map weighting (Spatial-η).

The MCFA technique generalizes the Cross Frame Attention (CFA) [20], as it enables the to-be-generated frame to attend to an arbitrary number of frames. We choose to attend to the first frame and the previous frame (as shown in lines $9-10$ of Alg. 1 and Fig. 3) to ensure long-range and short-range temporal consistency, respectively. MCFA intervenes in all the self-attention blocks of the SD UNet, by replacing the keys and values, that are originally computed from projections of the generating frame features, with the ones computed from the attended frames.

We also propose Spatial-η (line 12), that is a novel technique that enables to choose, on a pixel-by-pixel basis, whether to use DDIM or DDPM as a sampling scheme. This enables the usage of DDPM in regions of the images where novel content should be created (for example, when a new part of an object is entering the scene), while using DDIM in the other regions to ensure consistency and determinism where the already-present content is just moving. Note that this spatial map η^f

Table 1: Quantitative results.

		Frame Consistency		Motion Consistency	
		T2V0 [20]	MOTIONCRAFT	T2V0 [20]	MOTIONCRAFT
Fluid	Dragons	0.9664	**0.9991**	0.6846	**0.9637**
	Melting Man	0.9463	**0.9566**	0.7817	**0.8252**
Rigid Body	Satellite Scan	0.9588	**0.9875**	0.2852	**0.9219**
	Revolving Earth	**0.9812**	0.9696	**0.7213**	0.6783
Agents	Birds	0.9765	**0.9968**	0.8973	**0.9385**
Average		0.9658 $\pm$ 0.01	**0.9819** $\pm$ 0.02	0.6740 $\pm$ 0.23	**0.8655** $\pm$ 0.12

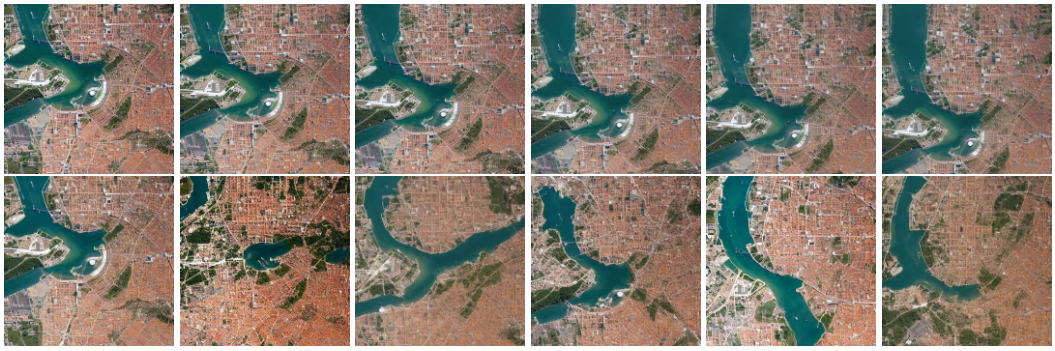

Figure 4: Rigid motion simulation: satellite orbit. Top: MOTIONCRAFT; Bottom: T2V0 [20].

can be obtained in multiple ways from the physical simulation. For example, η^f can be set to 1 in regions of the image where the flow is not well-defined (pointing outside of the image boundaries) or in regions where the optical flow field has discontinuities.

4 Experimental results

4.1 Experimental setting

In this section, we show examples of video generation based on different physics simulations: rigid body motion, fluid dynamics and multi-agent systems. Given an optical flow, we apply it on the SD latent space using MOTIONCRAFT (code is available at `https://mezzelfo.github.io/MotionCraft/`). Then, we compare our method to Text2Video-Zero [20] that, to the best of our knowledge, is the only diffusion-based zero-shot method for video generation.

We show qualitative results in Figs. 1, 4, 5, 6, 7, which we separately describe in the following sections. Table 1 reports two metrics to evaluate the quality of the generated videos. As done in previous works, we use the *Frame Consistency* metric, defined as the average cosine similarity of the CLIP embeddings of consecutive frames. However, this metric presents some limitations, as CLIP focuses on high-level semantic features and not on low-level details, resulting in high correlations even if the content changes but its semantics do not (as an example, see the video generated by T2V0 of the dragon in Fig. 6, which has a *Frame Consistency* of 0.97 even if the dragons are not the same dragons in each frame). To overcome some of these limitations, we propose a novel metric, named *Motion Consistency*, that measures how similar two frames are while accounting for the motion between them. We start from the observation that, if an object moves through the scene, its textures should remain almost the same, and, if we know its flow, we can bring back that object to overlap with its starting position. Then we can apply a similarity distance between the initial image and the next frame brought back by the reversed flow. Given two consecutive frames, we use a high-quality flow estimator (RAFT [32]) to estimate the optical flow between them and apply it on the second frame to reverse the motion. Then we compute the SSIM metric [34] on the first frame and the registered one.

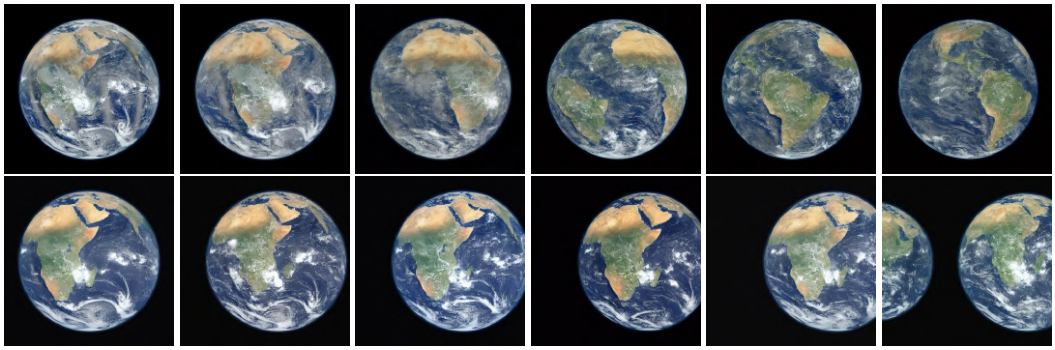

Figure 5: Rigid motion simulation: revolving Earth. Top: MOTIONCRAFT; Bottom: T2V0 [20].

Figure 6: Fluid simulation: dragon fire. Top: MOTIONCRAFT; Bottom: T2V0 [20].

4.2 Rigid Body flows

Fig. 4 shows a pivotal example where MOTIONCRAFT can be directly compared to the state-of-the-art T2V0, as in this case we use an optical flow equivalent to a their proposed shift along the vertical axis. This example shows a video generated starting from a satellite view of a city, and, by simulating the rectilinear motion of the satellite, new portions of the city appear from the top of the image. While T2V0 struggles with keeping temporal consistency, even with large structural elements (e.g., the river), MOTIONCRAFT is able to coherently scroll down the already present part of the city, while also generating new plausible content in the top part of the frames.

A similar case study is the Earth rotation in Fig. 5. Here, the optical flow is obtained by simulating a rotating sphere that was fitted to the first frame while keeping track of the starting and ending position of each point. As the Earth rotates, a slice disappears from one side and a new one needs to be generated on the opposite side. Thanks to the powerful natural image prior of SD, MOTIONCRAFT is able to autonomously generate other continents in the correct position, even if the text prompt contains no reference about them (see Appendix D for all the text prompts used in this paper). On the contrary, T2V0 is not able to rotate the Earth consistently while creating new content, as visible in the same Fig. 5.

4.3 Fluids dynamics

In this set of experiments, we use the Φ-flow [18] library to simulate fluid dynamics (by numerically solving Navier-Stokes equations) with the shape and position provided by the first frame I^0. Moreover, we can set up the simulation in different ways, depending on the numerical solvers, i.e. *Eulerian* (particle-based) or *Lagrangian* (grid-based), we can add rigid obstacles to the fluid or we can define a initial velocity and force fields. All these different options result in videos that can have the same starting frame but differ in their evolution according to the simulation constraints. We extract the velocity field of the simulation as a proxy for the optical flow. Examples of the velocity field can be seen in Appendix C.

Figure 7: Multi-agent system simulation: bird flock. Top: MOTIONCRAFT; Bottom: T2V0 [20].

Fig. 6 shows a fluid simulation of two dragons breathing fire. We can approximate the two initial fire breaths with two centered smoke balls, obtaining a binary mask that will be fed to the simulation. At this point, we run the simulation, solving the Navier-Stokes equations by sequentially evaluating advection, diffusion and pressure. The vorticity and the expansion of the smoke is due to the buoyancy force set in the desired direction, that in this case is such that the two balls cross near the middle of the image.

The figure shows that MOTIONCRAFT produces a consistent scene with a realistic animation of the fire breaths. Moreover, the global scene illumination seems to change accordingly, and a realistic occlusion of a dragon due to smoke gradually appears. This is mainly due to the MCFA mechanism, as we ablate in Sec. 4.5. In T2V0, the scene is not temporally consistent and shows increasingly more artefacts, such as color shifts or the fact that the right dragon changes with time, while the left one even disappears.

A similar analysis can be conducted for Fig. 1, where a simulation of a melting statue is shown. We can see that the generated video includes bouncing of parts on the ground before the fluid settles.

4.4 Multi-agent systems

Multi-agent systems are another interesting family of simulated dynamics. A simple agents model is the *Boids* model [24], consisting of a set of point-like agents (named boids) that move according to three steering behaviour rules: separation, as boids avoid collisions with nearby agents by steering away from them, alignment, as boids align their direction with that of nearby agents, and cohesion, as boids move towards the average position of nearby agents to stay together as a group. To simulate this system we used the agentpy [9] library, in which the number of agents, the simulation time-steps and different physical parameters related to the steering rules can be chosen.

An example is shown in Fig. 7, generating the temporal evolution of a flock of birds. As SD is not able to generate images with a controllable number of agents in specified positions, we start from an image where there is a single agent (a bird in the example). Then, we extract the corresponding latent vector patch with the attention map [8] related to the CLIP token containing the word "bird", and clone it to the simulated positions of the other agents. At this point, we evolve the frames according to the optical flow derived from the simulation velocity field.

While MOTIONCRAFT produces a realistic flock motion, T2V0 motion is not consistent and the number of birds changes in each frame.

4.5 Ablations

In this section, we ablate the contribution of the most important components/hyperparameters in the proposed pipeline. First, we start from investigating the impact of the cross-attention mechanism by comparing four different variants: i) each frame attends to itself (no MCFA); ii) each frame attends to the previous frame; iii) each frame attends to the first frame; iv) each frame attends to both the previous frame and the first frame (proposed MCFA). Visual results are shown in Fig. 8. As can be seen, the MCFA mechanism is necessary to generate plausible frames; moreover, attending only to the first frame reduces the overall motion, (e.g., always showing Africa as in the first frame), while

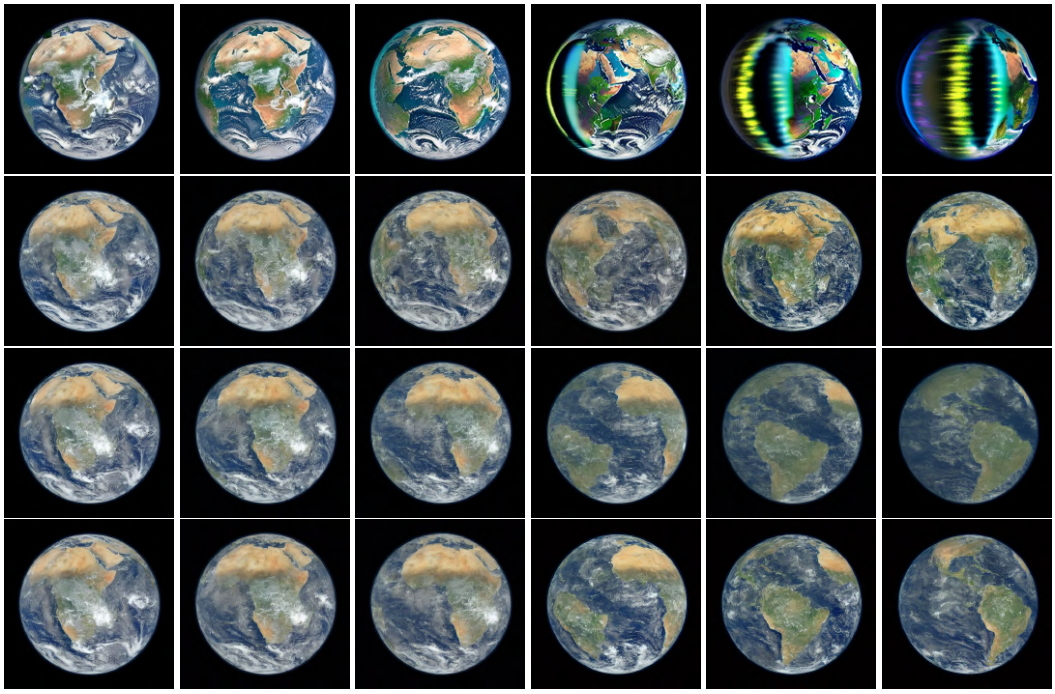

Figure 8: Ablation - Cross-Frame attention. First row: no cross frame attention; Second Row: Attend only to the initial frame; Third Row: Attend only to the previous frame; Fourth Row: Attend to the initial and preceding frame (ours).

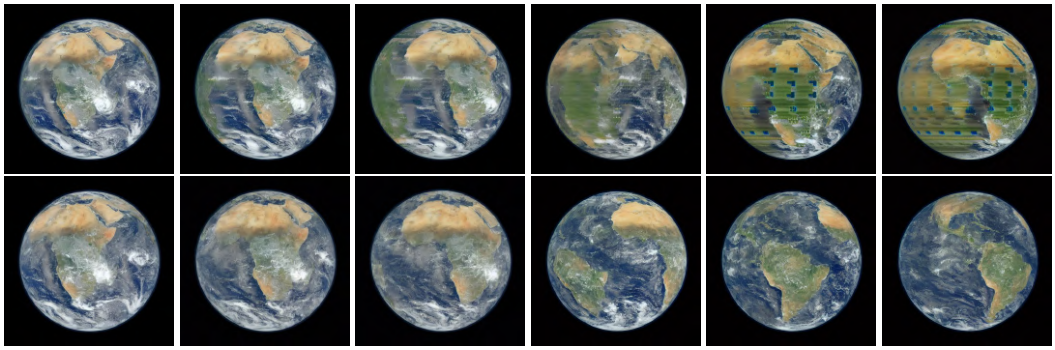

Figure 9: Ablation - Spatial-η. First Row: $\eta = 0$; Second Row: Spatial-η on.

only attending to the previous frame reduces color consistency. Overall, we demonstrate that the proposed MCFA, attending to both the first and the previous frame, represents the optimal solution to keep global consistency with the initial image and local consistency with the preceding frame.

Fig. 9 shows the ablation of the Spatial-η weighting technique. As shown, being able to sample with DDPM in some parts of the image is crucial in order to generate novel plausible content. Indeed, DDPM adds, during each reverse diffusion step, random white noise to the latent. We suppose that this allows to better sample from the real distribution, avoiding artefacts other components of the method, such as the warping operator or the MCFA, would otherwise introduce.

Finally, we ablated the partial inversion process, i.e., lines 2-6 in Alg. 1. Without the DDIM inversion, textures and details generated by SD cannot be brought into the next frame, resulting in corrupted videos. Visual results can be found in the Appendix B.3.

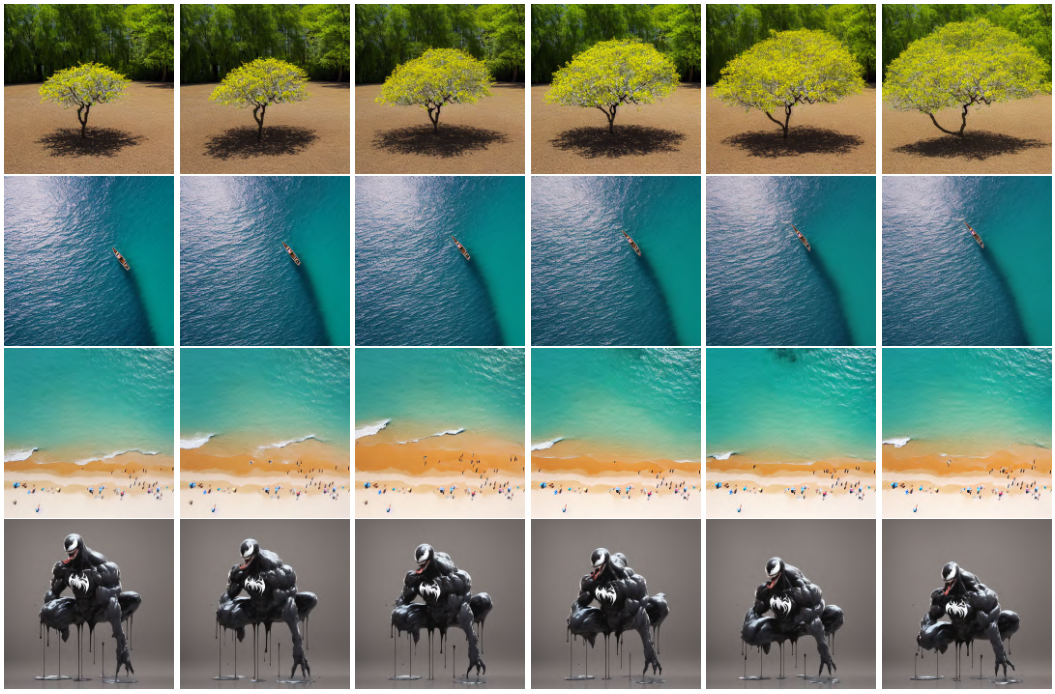

Figure 10: Additional videos from MOTIONCRAFT. The last row is obtained by applying MOTIONCRAFT to SDXL [23]

4.6 Additional Qualitative Results

Fig. 10 shows some additional results of MOTIONCRAFT. The first row shows a tree growing. This video was obtained using a simple constant outward-facing radial optical flow applied only on the foliage. Note that while the tree grows, its shadow evolution is coherent. As the input flow is zero in this part of the image, the shadow consistency is recovered only by Stable Diffusion. The last row show a video obtained by applying MOTIONCRAFT to SDXL. This shows the generalizability of MOTIONCRAFT to different diffusion models, with also different resolutions. Hence, MOTIONCRAFT is able to produce high-resolution videos with a high level of detail.

5 Conclusions

In this work, we have presented MOTIONCRAFT, a novel zero-shot approach for video generation. Our method allows to generate realistic videos with the image prior of Stable Diffusion and a physically-derived optical flow, without any additional training. MOTIONCRAFT warps the noise latent space according to the prescribed flow, and with a modified sampling process exploiting multi-frame cross-attention and the spatial-η variable sampling scheme generates novel plausible contents following the prescribed motion and tempoirally consistent. For the evaluations of the results, we relied on a standard metric and a proposed one, showing that our method is not only qualitatively but also quantitatively superior to the state-of-the-art of zero-shot video generation.

Acknowledgements

This publication is part of the project PNRR-NGEU which has received funding from the MUR - DM 352/2022. This work was partially supported by the European Union under the Italian National Recovery and Resilience Plan (NRRP) of NextGenerationEU, partnership on "Telecommunications of the Future" (PE00000001 - program "RESTART").

References

[1] Yuxiang Bao, Di Qiu, Guoliang Kang, Baochang Zhang, Bo Jin, Kaiye Wang, and Pengfei Yan. Latentwarp: Consistent diffusion latents for zero-shot video-to-video translation. *arXiv preprint arXiv:2311.00353*, 2023.

[2] Rishika Bhagwatkar, Saketh Bachu, Khurshed Fitter, Akshay Kulkarni, and Shital Chiddarwar. A review of video generation approaches. In *2020 International Conference on Power, Instrumentation, Control and Computing (PICC)*, pages 1–5, 2020. doi: 10.1109/PICC51425.2020.9362485.

[3] Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22563–22575, 2023.

[4] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18392–18402, 2023.

[5] Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, Clarence Ng, Ricky Wang, and Aditya Ramesh. Video generation models as world simulators. 2024. URL `https://openai.com/research/video-generation-models-as-world-simulators`.

[6] Shengqu Cai, Duygu Ceylan, Matheus Gadelha, Chun-Hao Paul Huang, Tuanfeng Yang Wang, and Gordon Wetzstein. Generative rendering: Controllable 4d-guided video generation with 2d diffusion models. *arXiv preprint arXiv:2312.01409*, 2023.

[7] Duygu Ceylan, Chun-Hao P Huang, and Niloy J Mitra. Pix2video: Video editing using image diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 23206–23217, 2023.

[8] Dave Epstein, Allan Jabri, Ben Poole, Alexei Efros, and Aleksander Holynski. Diffusion self-guidance for controllable image generation. *Advances in Neural Information Processing Systems*, 36:16222–16239, 2023.

[9] Joël Foramitti. Agentpy: A package for agent-based modeling in python. *Journal of Open Source Software*, 6(62):3065, 2021.

[10] Daniel Geng and Andrew Owens. Motion guidance: Diffusion-based image editing with differentiable motion estimators. In *The Twelfth International Conference on Learning Representations*, 2024. URL `https://openreview.net/forum?id=WIAO4vbnNV`.

[11] Michal Geyer, Omer Bar-Tal, Shai Bagon, and Tali Dekel. Tokenflow: Consistent diffusion features for consistent video editing. In *The Twelfth International Conference on Learning Representations*, 2024. URL `https://openreview.net/forum?id=lKK50q2MtV`.

[12] MSU Graphics and Media Lab. Msu video frame interpolation benchmark dataset, 2022. URL `https://videoprocessing.ai/benchmarks/video-frame-interpolation-dataset.html`.

[13] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-or. Prompt-to-prompt image editing with cross-attention control. In *The Eleventh International Conference on Learning Representations*, 2023. URL `https://openreview.net/forum?id=_CDixzkzeyb`.

[14] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021. URL https://openreview.net/forum?id=qw8AKxfYbI.

[15] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

[16] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, et al. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022.

[17] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *Advances in Neural Information Processing Systems*, 35:8633–8646, 2022.

[18] Philipp Holl, Nils Thuerey, and Vladlen Koltun. Learning to control pdes with differentiable physics. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=HyeSin4FPB.

[19] Itseez. Open source computer vision library. https://github.com/itseez/opencv, 2015.

[20] Levon Khachatryan, Andranik Movsisyan, Vahram Tadevosyan, Roberto Henschel, Zhangyang Wang, Shant Navasardyan, and Humphrey Shi. Text2video-zero: Text-to-image diffusion models are zero-shot video generators. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15954–15964, 2023.

[21] Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6038–6047, 2023.

[22] Haomiao Ni, Changhao Shi, Kai Li, Sharon X Huang, and Martin Renqiang Min. Conditional image-to-video generation with latent flow diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18444–18455, 2023.

[23] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023.

[24] Craig W Reynolds. Flocks, herds and schools: A distributed behavioral model. In *Proceedings of the 14th annual conference on Computer graphics and interactive techniques*, pages 25–34, 1987.

[25] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022.

[26] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pages 234–241. Springer, 2015.

[27] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, Devi Parikh, Sonal Gupta, and Yaniv Taigman. Make-a-video: Text-to-video generation without text-video data. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=nJfylDvgzlq.

[28] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, pages 2256–2265. PMLR, 2015.

[29] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising Diffusion Implicit Models. In *International Conference on Learning Representations*, 2020.

[30] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in Neural Information Processing Systems*, 32, 2019.

[31] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2020.

[32] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 402–419. Springer, 2020.

[33] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017.

[34] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4): 600–612, 2004.

[35] Jay Zhangjie Wu, Yixiao Ge, Xintao Wang, Stan Weixian Lei, Yuchao Gu, Yufei Shi, Wynne Hsu, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7623–7633, 2023.

[36] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023.

A Background

DDMs (Denoising Diffusion Models) [15, 28, 30] represents a generative modeling approach that leverage a noise diffusion process to model a data distribution starting from random noise. These models are based on a predefined Markovian forward noising chain that progressively adds Gaussian noise to the data $\boldsymbol{x}_0$ in an iterative procedure of T steps. The reverse diffusion process traverses back the Markov Chain and can be written as:

$$p_\theta(\boldsymbol{x}_{0:T}) = p(\boldsymbol{x}_T)\prod_{t=1}^{T} p_\theta(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t) \qquad p_\theta(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t) = \mathcal{N}(\boldsymbol{x}_{t-1} \mid \mu_\theta(\boldsymbol{x}_t, t), \sigma_t^2 \boldsymbol{I}) \; . \tag{1}$$

The training phase optimizes the parameters of the reverse process p_θ maximising an evidence lower bound (ELBO) over the target data. The work of [29] shows that is possible to construct a non-Markovian process defining a faster sampler (DDIM) that is compatible with the pretrained model. So starting from $p_\theta(\boldsymbol{x}_{0:T})$, it is possible to sample $\boldsymbol{x}_{t-1}$ using:

$$x_{t-1} = \sqrt{\alpha_{t-1}}\left(\frac{x_t - \sqrt{1-\alpha_t}\hat{\epsilon}_t}{\sqrt{\alpha_t}}\right) + \sqrt{1-\alpha_{t-1}-\sigma_t(\eta)^2}\cdot\hat{\epsilon}_t + \sigma_t(\eta)\varepsilon_t \tag{2}$$

where $\sigma_t(\eta) = \eta\sqrt{\frac{1-\alpha_{t-1}}{1-\alpha_t}}\sqrt{\frac{1-\alpha_t}{\alpha_{t-1}}}$ and $\eta \in (0, 1)$ is a parameter controlling the forward process, when $\eta = 0$, the sampling becomes deterministic, when $\eta = 1$, the process result in DDPM sampling. $\hat{\epsilon}_t$ is the estimated noise present in x_t, typically estimated with a UNet architecture [26]: $\epsilon_t(\cdot)$. Finally, ε_t is an independent normal stochastic variable. In this work we employ a Latent Diffusion Model [25] that perform the diffusion process over a compressed latent space, reducing the computational burden of training in pixel space, while keeping high perceptual quality. Before the diffusion process, a VQ-VAE [33] is trained; the input image is then encoded by the VQ-VAE Encoder $\mathcal{E}$ that reduces the spatial dimension. The generated features are decoded back to the image space when generating images by means of th VQ-VAE Decoder $\mathcal{D}$. The UNet architecture is tipically composed by convolutional layers followed by spatial self-attention layers and cross-attention conditioning layers. Recent works [4, 13, 20] propose to reprogram this mechanism to enhance consistency between frames by letting the currently generated frame to attend to the first frame by swapping the original attention keys (K) and values (V) with the keys and values of the first frame, leading to the Cross-Frame Attention (CFA) mechanism:

$$\text{Cross-Frame-Attn(Q,K,V)} = \text{Softmax}\left(\frac{Q^f \cdot K^1}{\sqrt{d_k}}\right)V^1 \tag{3}$$

where V^1 and K^1 represent the keys and values of the first frame, while Q^f represents the queries of the current frame, and d_k is the channel dimension of the keys. In this work we will use the notation $\epsilon_t(z, \mathcal{P}; \{a, b, c, \dots\})$, where z is a latent, $\mathcal{P}$ is the prompt, and $\{a, b, c, \dots\}$ is a *list* of latents to attend to, as MCFA enables to attends to a list of latents and not only to a single one.

Classifier-Free Guidance (CFG) [14] is a widely used technique to guide conditional generation process using a linear combination of conditional and unconditonal estimated scores:

$$\hat{\epsilon} = \epsilon_t(z, \mathcal{P}_\emptyset, \{\dots\}) + \gamma\left[\epsilon_t(z, \mathcal{P}, \{\dots\}) - \epsilon_t(z, \mathcal{P}_\emptyset, \{\dots\})\right] \tag{4}$$

where γ is the scaling factor, $\mathcal{P}_\emptyset$ represents the null condition and $\mathcal{P}$ is the target text prompt.

B Extendend Ablation Study

In this section we show the remaining ablations for the scene *Earth*, and additional ablations on two new scenes: *Dragons* and *Satellite Scan*. The ablations for cross frame attention mechanism can be found in Figures 11 and 12. The ablations of the Spatial-η are shown in Figures 13 and 14. Moreover, we also show the contribution of the inversion mechanism in Figures 15, 16, and 17.

B.1 Multiple Cross-Frame Attention Mechanism Ablation

Figure 11: Ablation - Cross-Frame attention. First row: no cross frame attention; Second Row: Attend only to the initial frame; Third Row: Attend only to the previous frame; Fourth Row: Attend to the initial and preceding frame (ours).

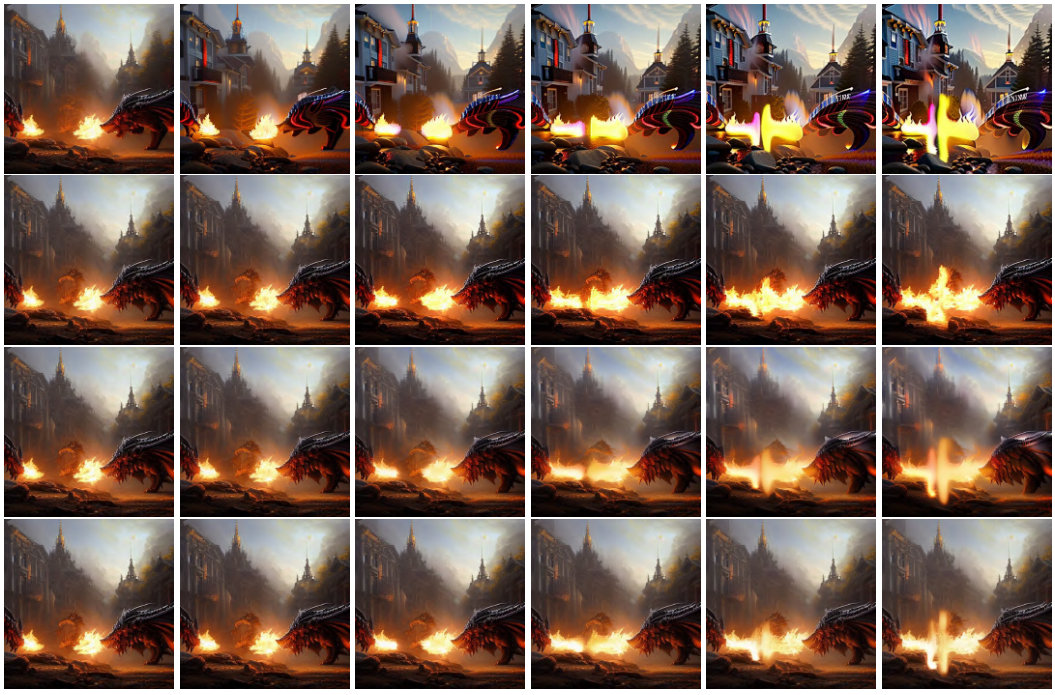

Figure 12: Ablation - Cross-Frame attention. First row: no cross frame attention; Second Row: Attend only to the initial frame; Third Row: Attend only to the previous frame; Fourth Row: Attend to the initial and preceding frame (ours).

B.2 Spatial eta

Figure 13: Ablation - Spatial-η. First Row: Spatial-η on; Second Row: $\eta = 0$.

Figure 14: Ablation - Spatial-η. First Row: Spatial-η on; Second Row: $\eta = 0$.

B.3 Inversion

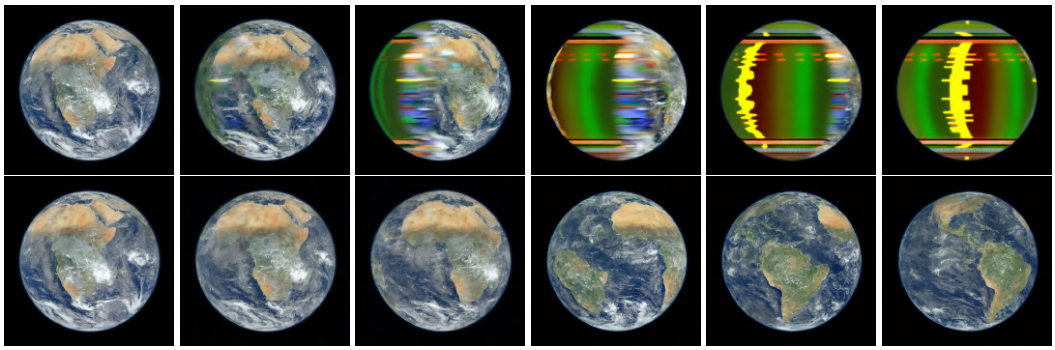

Figure 15: Ablation - Inversion Mechanism. First Row: Without Inversion; Second Row: With Inversion

Figure 16: Ablation - Inversion Mechanism. First Row: Without Inversion; Second Row: With Inversion

Figure 17: Ablation - Inversion Mechanism. First Row: Without Inversion; Second Row: With Inversion

C Obstacles, Different Physics and Additional Visual Results

In this section we showcase additional visual results of our method; all the generated videos can be found in the *Supplementary Material*. In Fig. 18 we show an example of a poured glass with MOTIONCRAFT (third row) and applying the same flow in image space (fourth row). In the first two rows of Fig. 18 we show the results of the Φ-flow physics simulator. Note that we simulate both the fluid as a set of particles (*Eulerian* simulations) in a specific position (blue balls in the first row of the figure) and two obstacles (orange objects) representing the glass and the jug. The corresponding optical flow that we used in MOTIONCRAFT is visualized in the second row of the figure.

As it can be seen, the optical flow applied to the image space produces some artefacts, such as deformations of the glass and the smoothness of the liquid due to the stretching of the pixels. On the other hand, when the same flow is applied to the noisy latent space through our method, the resulted video appears more realistic, avoiding such deformations.

Since Φ-flow is able to adopt both *Eulerian* and *Lagrangian* numerical solvers, we show the corresponding videos in Fig. 19 (second and fourth row). While the former decomposes the fluid in a set of particles, the latter models the fluid in the entire space as a fluid field. In both cases we extract from the simulation the (eventually extrapolated) velocity field (first and third row in Fig. 19) and we use it as the optical flow in the latent space, resulting in two different videos.

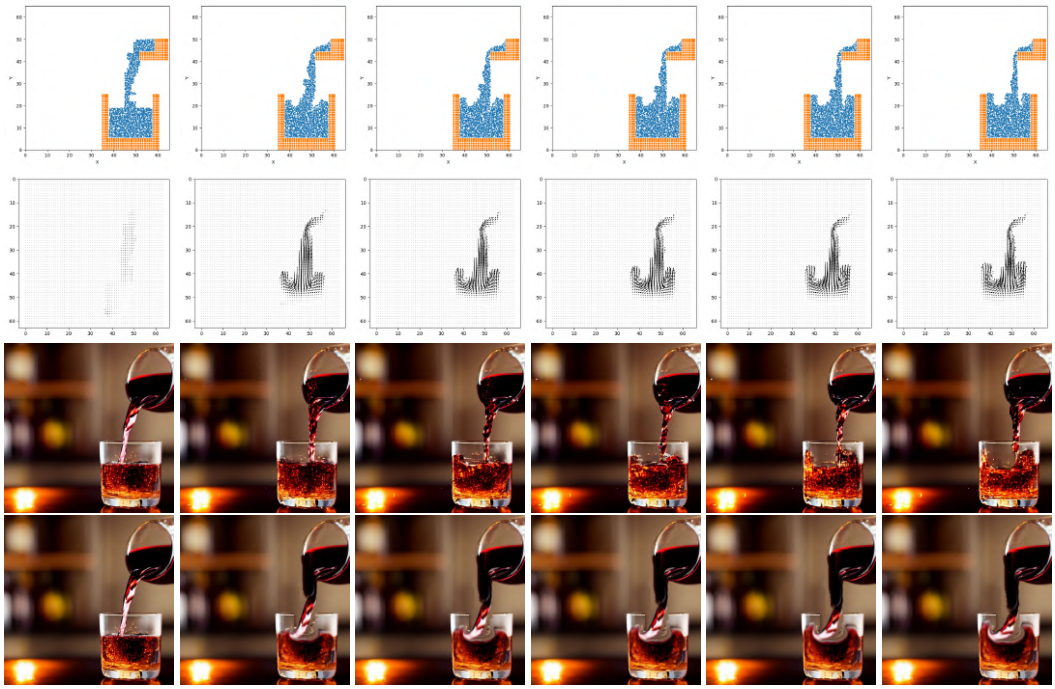

Figure 18: Fluid simulation: pouring drink. First row: Eulerian simulation performed with Φ-flow; Second row: resulting optical flow; Third row: MOTIONCRAFT; Fourth row: resulting video when optical flow is applied directly to pixel-space.

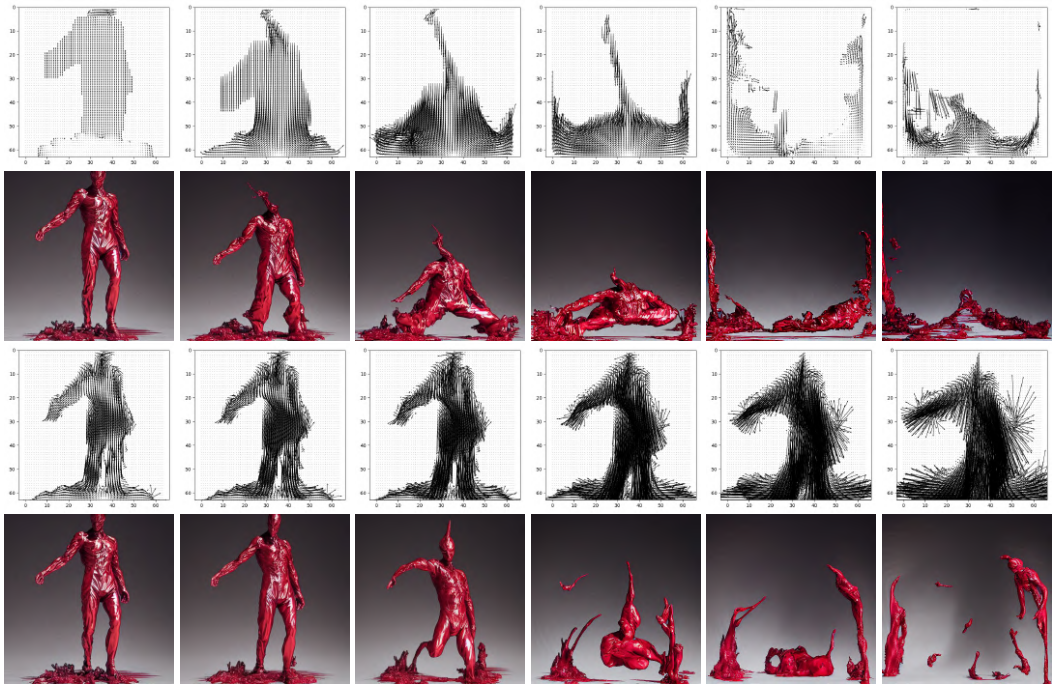

Figure 19: Smoke simulation: Evaporating man. First and second row: Optical flow and video generated by MOTIONCRAFT with an Eulerian simulation. Third and fourth row: Optical flow and video generated by MOTIONCRAFT with a Lagrangian simulation.

D Text Prompts

In this section we state the text prompts used in the generated videos for our method and T2V0. Note that while MOTIONCRAFT is able to start from a real or generated image (with almost zero error for the real image reconstruction), T2V0 needs a hyper-parameters tuning due to a high guidance scale (not supporting direct inversion of real images).

- *Fighting Dragons*: "Two dragons fighting while breathing fires to each other. The flames are blazing and majestic light. Theatrical, character concept art by ruan jia, thomas kinkade, and trending on Artstation."
- *Melting Man* (both versions): "transparent man made by water and smoke, in style of Yoji Shinkawa and Hyung-tae Kim, trending on ArtStation, dark fantasy, great composition, concept art, highly human made of water and foam, in the style of Pierre Koenig, red pigment, pastel paint, pink color scheme"
- *Satellite Scan*: "a satellite image of a city"
- *Revolving Earth*: "a close up of a picture of the earth from space."
- *Flock of birds*: "a small flock bird flying in the sky at the sunset"
- *Pouring drink*: "wine falling on a empty glass"

For the text prompts of *Fighting Dragons* and *Melting Man* we leveraged MagicPrompt (for which we credit Gustavo Santana), a tool for rewriting simple text prompts to create more appealing starting images with Stable Diffusion.

For each example, the negative prompt $\mathcal{P}_\emptyset$ is equal to "poorly drawn, cartoon, 2d, disfigured, bad art, deformed, poorly drawn, extra limbs, close up, b&w, weird colors, blurry"

E Limitations and future work

In this section we discuss the limitations of the proposed approach. Being a zero-shot approach, MOTIONCRAFT relies on the pretrained text-to-image model, i.e. Stable Diffusion, and it can inherit some limitations from it, such as not exact DDIM inversion. Hence, by exploiting other diffusion models we could improve our method as well.
Experimentally, we observed a global color shift, getting stronger in the last frames of the generated videos. We noted that the proposed MCFA strategy partially solved this, but a better solution could be attending to all the previous generated frames (albeit resulting in a memory and run-time complexity increase). Moreover, MOTIONCRAFT depends on the optical flow derived from physics simulations but there are some dynamics that may be difficult to simulate (e.g. the motion of a dancer), thus limiting the generality of the generated videos. However, we speculate that it might be possible to devise a generative model of optical flows conditioned on a starting frames and a prompt, while also being constrained by a physics simulator. This could readily provide inputs to MOTIONCRAFT and have the advantage of disentangling learning of motion from learning of content. A future direction could also employ a better interaction between the image generator and the physics simulator, in order to have a closed feedback-loop framework leading to more physical fidelity in the generated frames. In this work we have shown videos generated by different physical simulations, but as future work we could also combine them to generate more complex scenes with different physics mixed together.

F Implementation Details and Licenses

We used the following hyperparameters throughout the work if not explicitely said otherwise. We set $\tau = 400$, the number of inference steps (both for DDIM inversion and for inverse diffusion) is set to 200 and the used model is `runwayml/stable-diffusion-v1-5` (license CreativeML Open RAIL-M). All our experiments are done on a single NVIDIA A6000 (48GB); video generation runs in minutes (1-5min) on a single GPU. Our provided code is available under MIT license. The *Earth* image is a composite of six separate orbits taken on January 23, 2012 by the Suomi National Polar-orbiting Partnership satellite (Credit: NASA/NOAA).

G Broader Impact

Synthetic video generation is a powerful technology that can be misused to create fake videos, hence it is important to limit and safely deploy these models. From a safety perspective, we emphasize that MOTIONCRAFT does not add any new restrictions nor does it relax any existing ones with respect to our base text-to-image model. Moreover MOTIONCRAFT, using existing text-to-image diffusion models, does not need extra training or adjustments. This means we avoid the large environmental costs associated with training new models. One possible broader impact of MOTIONCRAFT is its usage by scientists across various fields to visualize their simulations, thereby offering AI-based visualization of physical processes to a wider scientific audience.

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: All the claims in the abstract and introduction sections are supported by experimental evidence throughout the paper and reflect the results of our method.

   Guidelines:

   - The answer NA 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 NA 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 proposed work in the Appendix Section E.

   Guidelines:

   - The answer NA 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: [NA]

Justification:

Guidelines:

- The answer NA 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: We report a detailed pseudocode 1 from which it is possible to reproduce the main experimental results of our work. The method with all hyperparameters and implementation details are present in the Appendix F.

   Guidelines:

   - The answer NA 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: There is a link in the abstract pointing to the official project page, containing a public version of this code.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.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://nips.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, etc.) necessary to understand the results?

Answer: [Yes]

Justification: We specify the experimental setting in section 4.1 and in the Appendix F

Guidelines:

- The answer NA 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 report mean and std of the results in Table 1.

Guidelines:

- The answer NA 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: We specify the compute resource and time of execution in Appendix F

   Guidelines:

   - The answer NA 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: The work respects the Code of Ethics.

   Guidelines:

   - The answer NA 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: [Yes]

    Justification: We discuss the broader impact of the work in Appendix G

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA 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., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification:

    Guidelines:

    - The answer NA 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 all the used assets throughout the work and we report the correct version used for our experiments.

    Guidelines:

    - The answer NA 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 release the code along with the documentation needed to run the experiments.

    Guidelines:

    - The answer NA 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: [NA]

    Justification:

    Guidelines:

    - The answer NA 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: [NA]

    Justification:

    Guidelines:

    - The answer NA 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.