WorldReel: 4D Video Generation with Consistent Geometry and Motion Modeling

Supplementary Material

1. Method

1.1. Temporal DPT Architecture

As described in Sec. 3.3, we employ a customized temporal DPT model to predict the unified 4D representations (depth, point cloud, camera, scene flow, and motion mask) from the input geo-motion latent. The architecture is inspired by DPT [42] architecture and adapted for multi-task video prediction (see Figure 6).

First, a pyramid encoder with sequential 3D convolutions extracts multi-scale dense features from the input geo-motion latent. We then employ a DPT-style fusion backbone to process and aggregate these multi-scale features. Notably, several temporal layers utilizing temporal transformer models are inserted into the DPT architecture to effectively model spatiotemporal information. This backbone produces a single unified spatio-temporal feature map, which is then upscaled to the target resolution. From this unified feature, we use separate, lightweight output heads to predict the different geometry and motion tasks in parallel. The camera branch is treated differently: we apply average pooling to the unified feature map to obtain a per-frame feature vector. This vector is then processed by an additional temporal layer to explicitly model the camera trajectory, which requires strong temporal consistency.

1.2. Pseudo 3D Motion Label

We introduce our pipeline to produce 2D/3D motion labels for explicit scene dynamics modeling in Sec. 3.4; here, we provide the details of filtering noisy labels.

Given the forward and backward optical flow $F^{\text{2d}}_{i\to i+1}, F^{\text{2d}}_{i+1\to i} \in \mathbb{R}^{H\times W\times 2}$, and their per-pixel uncertainties $\sigma^{\text{2d}}_{i\to i+1}$, $\sigma^{\text{2d}}_{i+1\to i} \in \mathbb{R}^{H\times W}$, for pixel $\mathbf{u}=(x,y)$ in frame i, we can get the forward-mapped pixel $\mathbf{q}(\mathbf{u}) = \mathbf{u} + F^{\text{2d}}_{i\to i+1}(\mathbf{u})$. The pseudo scene flow label $\hat{F}^{3d}_i(u) = P_{i+1}(\mathbf{q}(\mathbf{u})) - P_i(\mathbf{u})$ may be noisy or invalid, so we use a validity mask M^{flow}_i to filter out potentially incorrect scene flow labels.

A pixel $\mathbf{u}$ is considered valid by intersecting multiple requirements: (i) Boundary check. The warped pixel $\mathbf{q}(\mathbf{u})$ must lie within the image boundaries $[0, W-1]\times[0, H-1]$, denoted as M^{bound}_i. (ii) Instance check. The pixels in two corresponding images should have the same instance/foreground labels, denoted as M^{inst}_i. For BEDLAM [4], Dynamic Replica [29] and OmniworldGame [80] that do not provide instance labels, we use the ground truth foreground mask to check the label consistency. Also, we utilize the ground truth instance labels retrieved from simulation for PointOdyssey [78] and the pseudo foreground instance labels produced by ViPE [22] for SpatialVid [58]. (iii) Flow uncertainties. The uncertainty estimates from SEA-RAFT [63] must be below a threshold τ_σ for both flows: $M^{\text{unc}}_i(\mathbf{u}) = \sigma^{\text{2d}}_{i\to i+1}(\mathbf{u}) < \tau_\sigma$ and $\sigma^{\text{2d}}_{i+1\to i}(\mathbf{q}(\mathbf{u})) < \tau_\sigma$. We set $\tau_\sigma = 3$ in implementation. (iv) Forward-backward consistency. The forward and backward 2d flows should be consistent. We check this by measuring the consistency error $E_{\text{cons}}(\mathbf{u}) = \left\|F^{2d}_{i\to i+1}(\mathbf{u}) + F^{2d}_{i+1\to i}(\mathbf{q}(\mathbf{u}))\right\|_2$. The consistency validity is checked by $M^{\text{cons}}_i(\mathbf{u}) = E_{\text{cons}}/(\left\|F^{2d}_{i\to i+1}(\mathbf{u}))\right\|_2 + (\left\|F^{2d}_{i+1\to i}(\mathbf{q}(\mathbf{u}))\right\|_2 + \epsilon) < \tau_{\text{cons}}$, where we set $\tau_{\text{cons}} = 0.2$.

The overall validity mask is

$$M^{\text{flow}}_i = M^{\text{bound}}_i \wedge M^{\text{inst}}_i \wedge M^{\text{unc}}_i \wedge M^{\text{cons}}_i \tag{6}$$

1.3. Training Loss

For $\mathcal{L}_{\text{flow}}$ in Eq. 2, we use L1 loss for sharp supervision, and use the $\hat{M}^{\text{bg}}_i$ and $\hat{M}^{\text{fg}}_i$ masks from $\hat{M}_i$ label to re-weight the flow supervision and emphasize foreground motion. For the dynamic (foreground) part of the 4D scene, we use the pseudo ground truth $\hat{F}^{\text{3d}}$ to supervise and apply the scene flow validity mask M^{flow}, $\mathcal{L}^{\text{dyn}}_{\text{flow}} = \Sigma_{i,\mathbf{u}} M^{\text{flow}}_i(\mathbf{u}) \left\|\hat{F}^{\text{3d}}_i - F^{\text{3d}}_i(\mathbf{u})\right\|_1 / \Sigma_{i,\mathbf{u}} M^{\text{flow}}_i(\mathbf{u})$. For the static (background) part of the 4D scene, the ground truth motion is $\mathbf{0}$, $\mathcal{L}^{\text{static}}_{\text{flow}} = \Sigma_{i,\mathbf{u}} \hat{M}^{\text{bg}}_i(\mathbf{u}) \left\|F^{\text{3d}}_i(\mathbf{u})\right\|_1 / \Sigma_{i,\mathbf{u}} \hat{M}^{\text{bg}}_i(\mathbf{u})$. The flow loss can be formulated as: $\mathcal{L}_{\text{flow}} = \mathcal{L}^{\text{dyn}}_{\text{flow}} + \alpha_{\text{static}} \mathcal{L}^{\text{static}}_{\text{flow}}$, with a small coefficient $\alpha_{\text{static}} = 0.1$.

2. Experiment

2.1. Evaluation Bench

To comprehensively evaluate the performance of our method on in-the-wild 4D scene generation, we utilize filtered videos from the SpatialVid [58] dataset. Specifically, we utilize real-world videos in group-0001 to group-0002[1] as the validation split. Based on this video collection, we construct the two evaluation benchmarks used in our experiments: the *general* motion set and *complex* motion set. The *general* motion set contains 500 videos that are randomly sampled. To select samples for the *complex* motion set, we utilize the pseudo 3D motion labels F^{3d} to calculate a 3D motion magnitude for each video clip, which only assesses the object dynamics

[1] https://huggingface.co/datasets/FelixYuan/SpatialVID-HQ

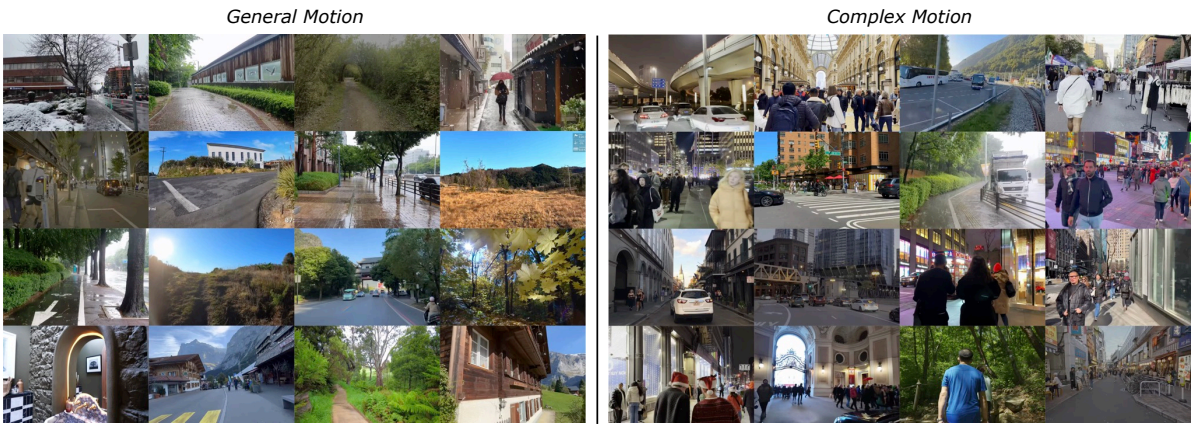

Figure 5. Condition image examples of *general* motion and *complex* motion sets in our evaluation bench.

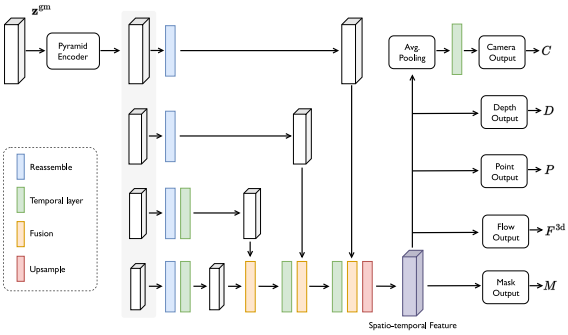

Figure 6. Overview of the temporal DPT model architecture.

in a scene. We gather the 500 samples with the highest 3D motion magnitude to get the *complex* set.

Figure 5 illustrates some examples of condition images in the *general* motion set and *complex* motion set. The *general* motion set primarily comprises videos depicting largely static scenes, some featuring minimal object motion such as sparse pedestrians or vehicles. This set encompasses a diverse range of scene categories, including city streets, forests, courtyards, and architectural environments. The dominant motion in these videos is attributed to camera ego-motion. In contrast, the *complex* motion set encompasses significantly more intricate scenarios, including both indoor and outdoor settings with diverse dynamic entities like pedestrians and vehicles. A key characteristic of this set is the presence of numerous and densely moving objects, presenting substantial challenges for coherent 4D scene generation.

2.2. Additional Quantitative Results

Depth Reprojection Consistency (DRC). To further evaluate geometry consistency without relying on pseudo ground truth, we report depth reprojection consistency (DRC). Specifically, we reproject the generated depth from each frame to all other frames using the generated camera parameters and measure the reprojection errors. We additionally report *DRC w/ mask*, which excludes pixels inside the predicted dynamic masks.

Table 4. Depth reprojection consistency (DRC). *DRC w/ mask* excludes pixels in dynamic masks.

method	DRC	DRC w/ mask
4DNeX	0.0643	-
GeoVideo	0.1081	-
w/o geomotion	0.1084	0.0930
w/o joint	0.0812	0.0584
WorldReel	**0.0381**	**0.0231**

Table 5. Inference runtime and memory usage on a single H200 GPU without CPU offloading.

	4DNeX	CogVideoX	**WorldReel**
Runtime (s)	725	108	113
Memory (GB)	57.4	26.3	44.8

Table 6. Pairwise user preference study.

	baseline	**Ours**
GeoVideo	13.0%	**87.0%**
DimensionX	4.3%	**95.7%**
4DNeX	15.0%	**85.0%**

Inference runtime. We report the inference runtime and memory usage of different methods in Table 5. All measurements are conducted on a single H200 GPU without CPU offloading. Compared with CogVideoX, WorldReel introduces only a small runtime overhead while producing explicit 4D outputs, and remains substantially more efficient than 4DNeX.

User study. We conduct a pairwise user preference study with 15 participants on 60 randomly sampled real validation images. Given the same input image and two generated videos, participants are asked to choose the better result overall according to (i) motion naturalness and physical plausibility, (ii) geometric consistency, and (iii) visual quality. As shown in Table 6, users strongly prefer our results over all baselines.

2.3. Qualitative Ablation

We present qualitative ablation comparisons in Figure 7. As demonstrated in both the sampled frames and the video, our full model generates the most realistic and temporally stable videos for in-the-wild scenes. Applying joint training with regularization on the RGB-only model ("w/o g.m.") clearly produces visual artifacts and ghosting, degrading the scene's stability. Similarly, removing the joint training stage and regularization ("w/o joint") exhibits reduced subject consistency, resulting in incoherent motion for dynamic objects and people. In contrast, the full model yields the smoothest and most coherent non-rigid human motion, resulting in natural and high-quality complex dynamics.

2.4. Qualitative Comparison

We present the full qualitative comparisons with DimensionX [52], 4DNeX [10], and GeoVideo [3] in Figure 8 corresponding to examples in Figure 3. Prior methods often exhibit geometry drift and motion inconsistencies (e.g., warped facades, misaligned vehicles), while our results better preserve scene layout and maintain coherent camera and non-rigid dynamics.

We show more generation results by our model (WorldReel) for static/dynamic in-the-wild scenes in Figure 9. Results demonstrate the strong generalization capability of our method. WorldReel successfully generates high-fidelity videos across a wide spectrum of scenarios, including diverse natural landscapes, urban traffic, and complex architectural environments at various scales. Aided by our geometry-motion augmented latent and targeted regularization, the generated videos exhibit robust geometric consistency. Notably, even in scenes with complex, non-rigid motion, such as multiple pedestrians, the generated dynamics remain coherent and smooth, underscoring our model's ability to accurately model a consistent underlying 4D scene.

2.5. Visualization

We provide generated 4D video visualization and generated 4D scene rendering videos. Please refer to the website for full videos.

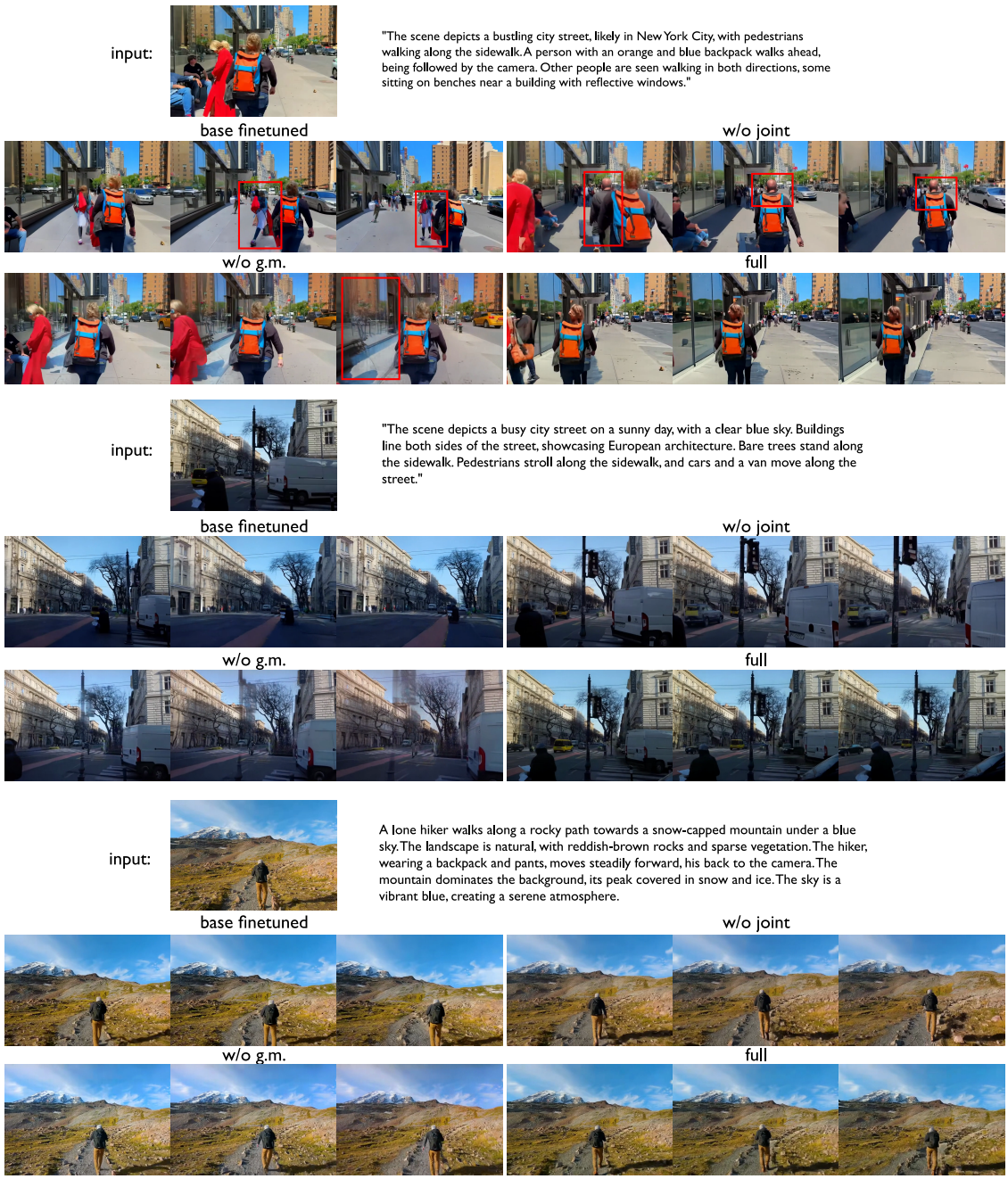

Figure 7. Qualitative ablation on video generation. We demonstrate the condition images, text prompts, and sampled frames (16/32/48-th) of the generated videos here. Please refer to the website for full videos.

Figure 8. Full examples of qualitative image-to-video comparison on in-the-wild scenes in Figure 3. We demonstrate the condition images, text prompts, and sampled frames (16/32/48-th) of the generated videos here. Please refer to the website for full videos.

Figure 9. Image-to-video generation results. We show more image-to-video generation on general in-the-wild scenes. Please refer to the website for full videos.