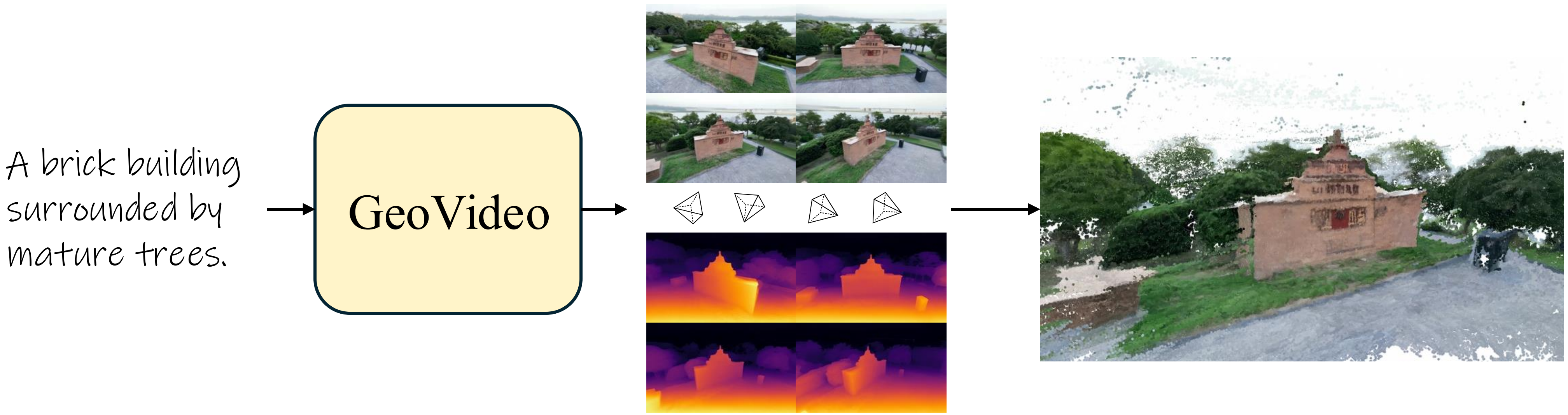
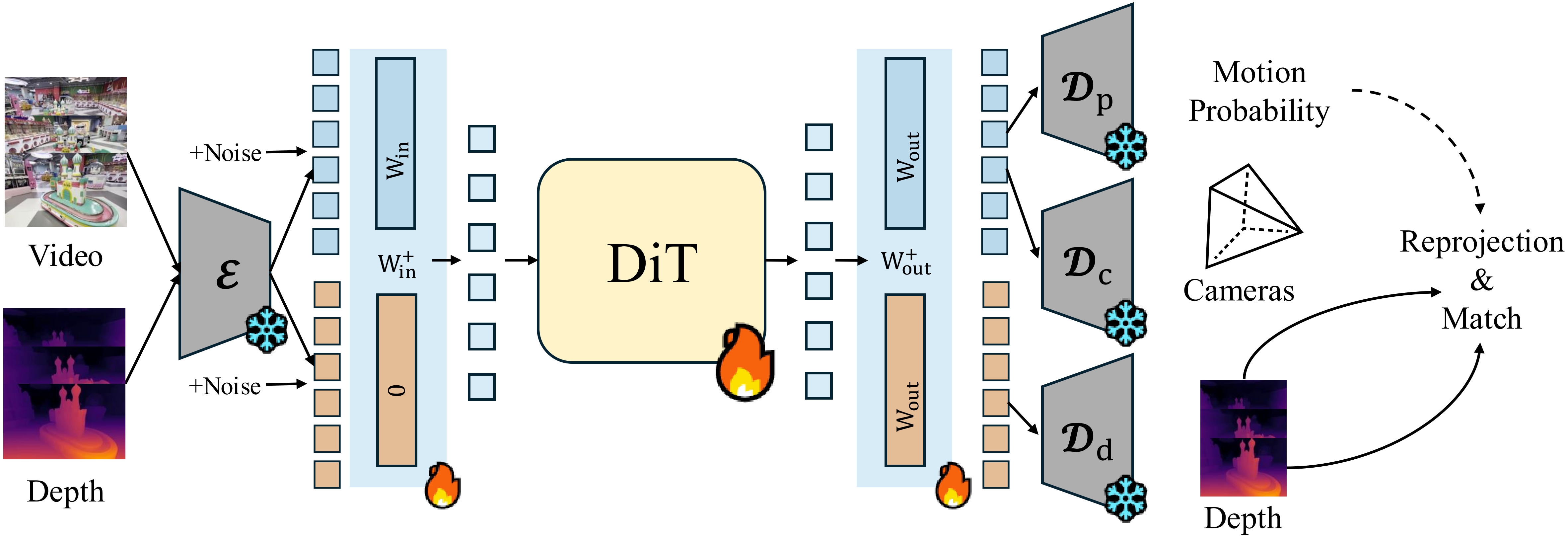
GeoVideo: Introducing geometric regularization into video generation model

Yunpeng Bai, Shaoheng Fang, Chaohui Yu, Fan Wang, Qixing Huang
[1] The University of Texas at Austin, [2] DAMO Academy, Alibaba Group, [3] Hupan Lab



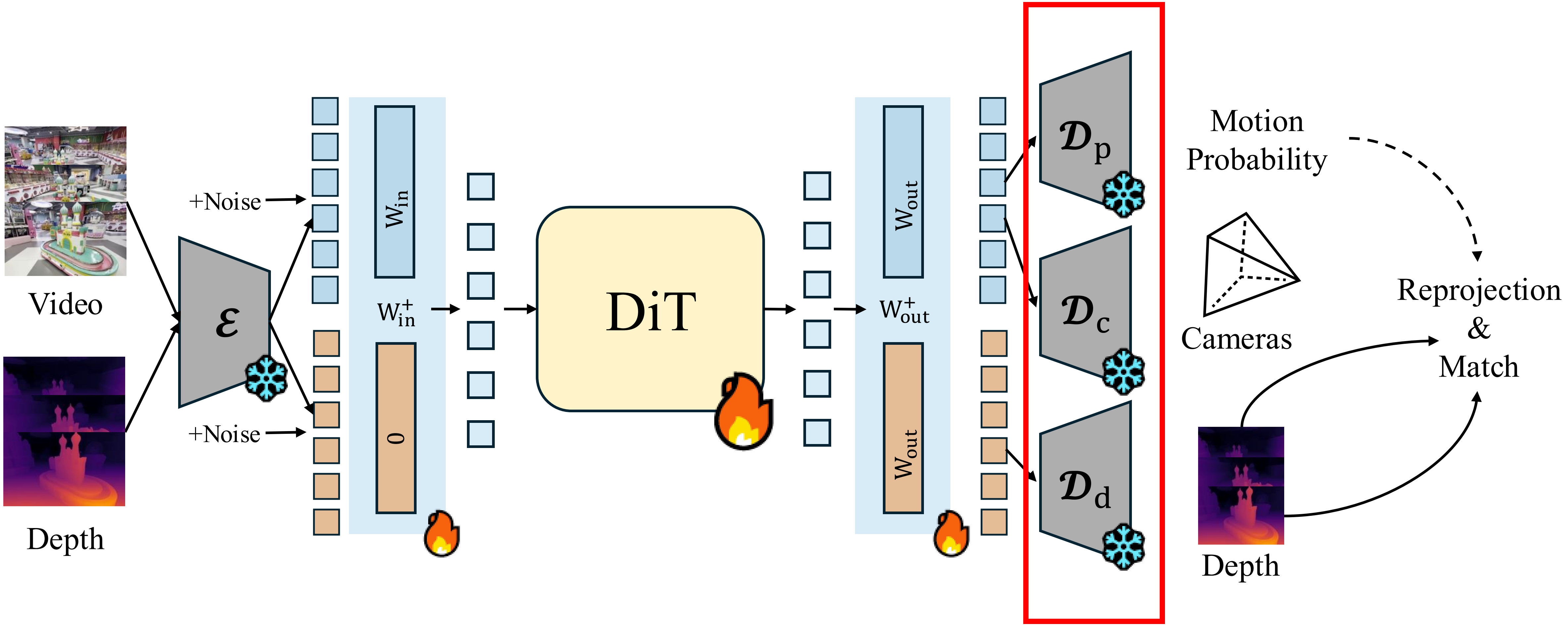
Why Geometry Matters in Video Generation

Core Idea: Add Depth as Geometric Supervision



$$\mathbf{z} = [\mathbf{z}^{\mathrm{RGB}}; \mathbf{z}^{\mathrm{D}}] = [E(\mathbf{x}_{1:T}^{\mathrm{RGB}}); E(\mathbf{x}_{1:T}^{\mathrm{D}})],$$

Core Idea: Add Depth as Geometric Supervision



Geometric Regularization Loss

$$\mathbf{X}_i = \mathbf{P}_i \cdot \pi^{-1}(\mathbf{D}_i, K),$$

where π^{-1} denotes backprojection from depth to 3D coordinates.

$$\mathcal{X}_{\text{global}} = \bigcup_{i=1}^{T} \mathbf{X}_i.$$

We denoise $\mathcal{X}_{\text{global}}$ using voxel grid downsampling and statistical outlier removal to improve robustness and computational efficiency.

$$\hat{\mathbf{D}}_i(\mathbf{u}) = \pi_z(\mathbf{P}_i^{-1} \cdot \mathbf{x}), \quad \mathbf{x} \in \mathcal{X}_{\text{global}},$$

$$\mathcal{L}_{\text{geo}} = \frac{1}{T} \sum_{i=1}^{T} \frac{1}{|\mathcal{V}_i|} \sum_{\mathbf{u} \in \mathcal{V}_i} \mathbb{1}(|\hat{\mathbf{D}}_i(\mathbf{u}) - \mathbf{D}_i(\mathbf{u})| < \delta) \cdot |\hat{\mathbf{D}}_i(\mathbf{u}) - \mathbf{D}_i(\mathbf{u})|,$$

where $\mathcal{V}_i$ is the set of valid pixels and δ is a tolerance threshold set to 0.05.

Two-Stage Training

$$W_{\text{in}}^{+} = \begin{bmatrix} W_{\text{in}} \\ \mathbf{0} \end{bmatrix} \in \mathbb{R}^{2C_v \times C_t}, \quad b_{\text{in}}^{+} = b_{\text{in}} \in \mathbb{R}^{C_t},$$
$$W_{\text{out}}^{+} = [W_{\text{out}} \quad W_{\text{out}}] \in \mathbb{R}^{C_t \times 2C_v}, \quad b_{\text{out}}^{+} = \begin{bmatrix} b_{\text{out}} \\ b_{\text{out}} \end{bmatrix} \in \mathbb{R}^{2C_v}.$$

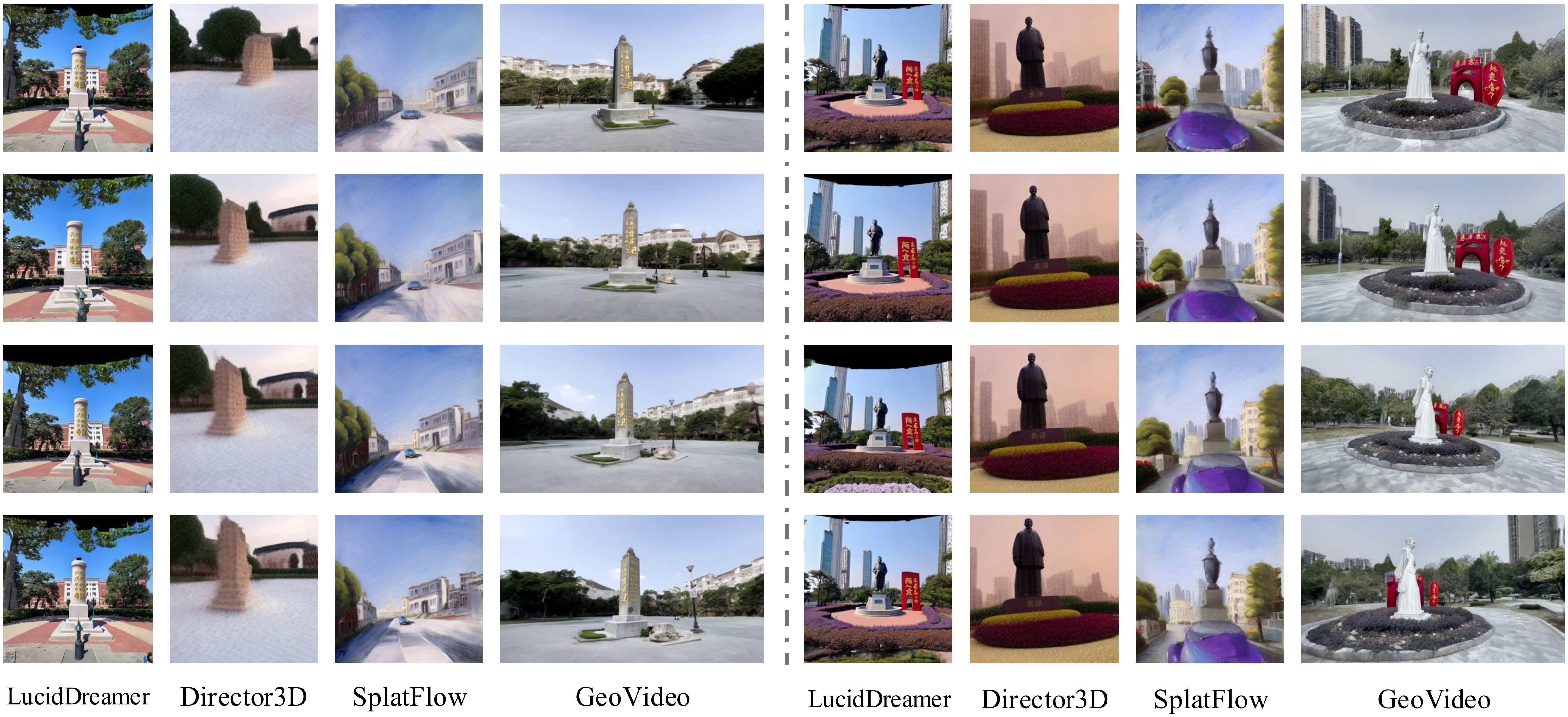
Stage 1: RGB-D Joint Generation.

$$\lambda_{\text{depth}}(t) = \min(1.0, 0.1 + \alpha t),$$
$$\mathcal{L}_{\text{stage-1}} = \mathcal{L}_{\text{diff}}^{\text{RGB}} + \lambda_{\text{depth}}(t) \cdot \mathcal{L}_{\text{diff}}^{\text{D}}.$$

Stage 2: Geometric Regularization.

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{diff}}^{\text{RGB}} + \lambda_{\text{depth}} \cdot \mathcal{L}_{\text{diff}}^{\text{D}} + \lambda_{\text{geo}} \cdot \mathcal{L}_{\text{geo}}.$$


CogVideoX
CogVideoX-tuned
GeoVideo
N/A
CogVideoX
CogVideoX-tuned
GeoVideo


LucidDreamer
Director3D
SplatFlow
GeoVideo
LucidDreamer
Director3D
SplatFlow
GeoVideo

Thank you !