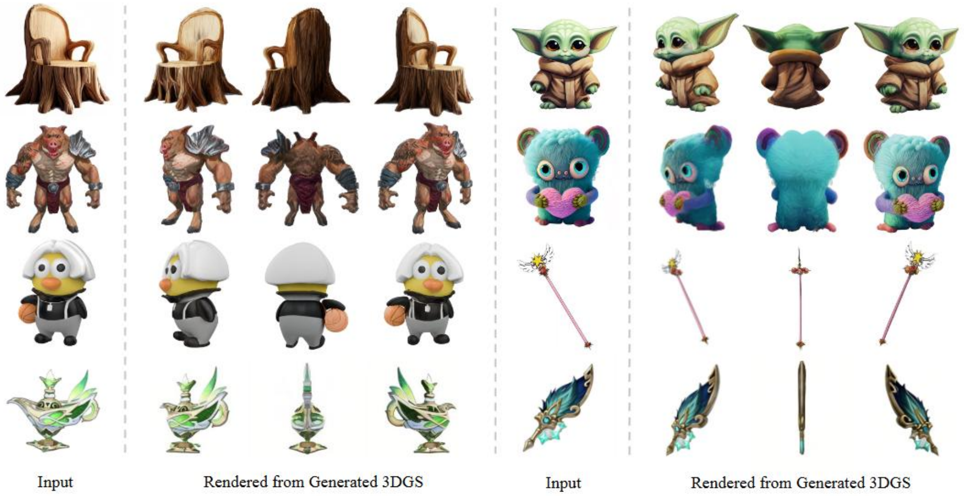
Ouroboros3D: Image-to-3D Generation via 3D-aware Recursive Diffusion

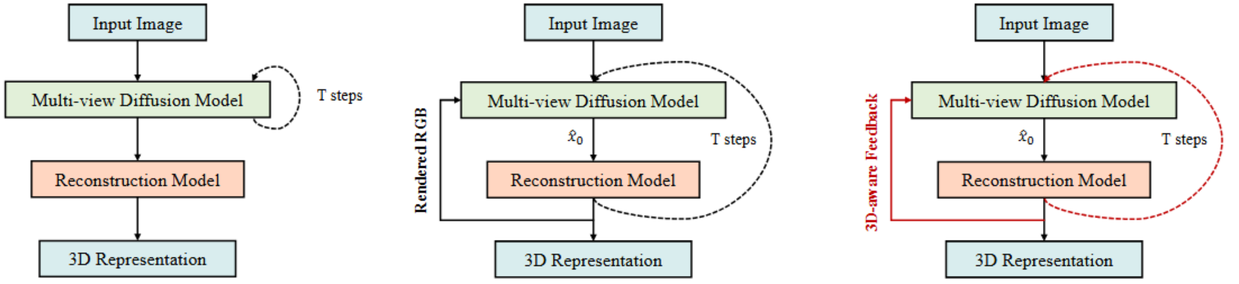
Hao Wen[1*], Zehuan Huang[1,3*], Yaohui Wang[2], Xinyuan Chen[2], Lu Sheng1✉
[1]Beihang University [2]Shanghai AI Laboratory [3]VAST

https://costwen.github.io/Ouroboros3D/



Motivation

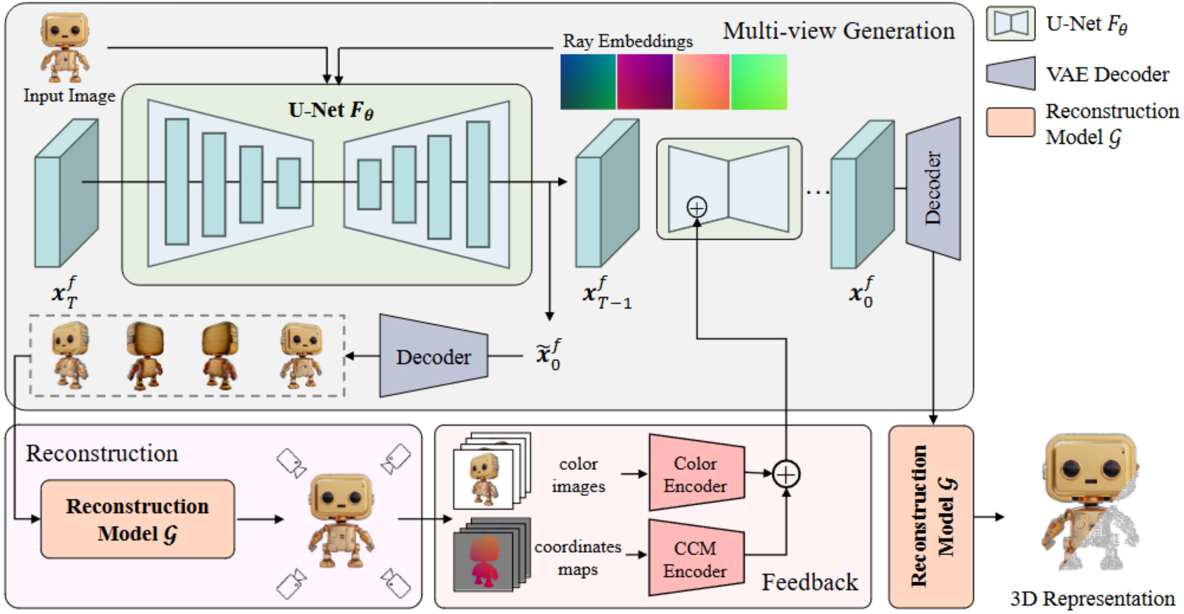
The motivation for the Ouroboros3D framework arises from separately generation and reconstruction model training. Ouroboros3D integrates multi-view generation and 3D reconstruction into a recursive diffusion process.



(a) Two-stage pipeline (Inference) (b) Iteration of two stages (Inference) (c) **Ouroboros** framework **(Training & Inference)**

Pipeline: Multi-view Diffusion Model

We introduce a self-conditioning mechanism, feeding the 3D-aware information obtained from the reconstruction module back to the multi-viewgeneration process. The 3D-aware recursive diffusion strategy iteratively refines the multi-view images and the 3d model.



Key Idea: 3D aware self condition feedback

Training+ Inference pseudocode

Algorithm 1 Training

Input: x, cond_image, cameras, timestep
Output: loss
// Returns the loss on a training example x. Details about EDM are omitted here.

```
begin
    noise ← Sample from Normal Distribution
    noisy_x ← Add_Noise(x, noise, timestep)
    pred_x ← F(noisy_x, cond_image, timestep, cameras)
    pred_i ← VAE_Decoder(pred_x)
    self_cond ← G(pred_i, cameras, timestep)
    if Random_Uniform(0, 1) > 0.5 then
        pred_x ← F(noisy_x, cond_image, timestep, cameras, self_cond)
    end
    loss_mv ← MSE_Loss(pred_x, x)
    loss_recon ← MSE_Loss(self_cond, x) + LPIPS_Loss(self_cond, x)
    loss ← loss_mv + loss_recon
    return loss
end
```

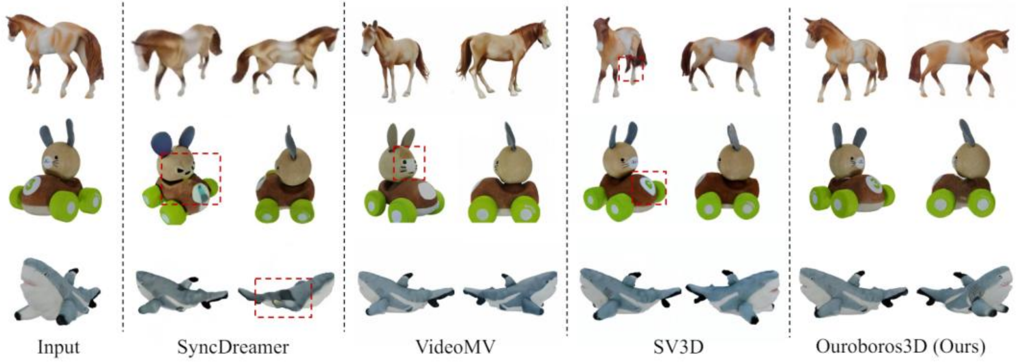
Algorithm 2 Inference

Input: cond_image, cameras, timesteps
Output: images, 3d_model
// Generate multi-view images and 3D model from a condition image.

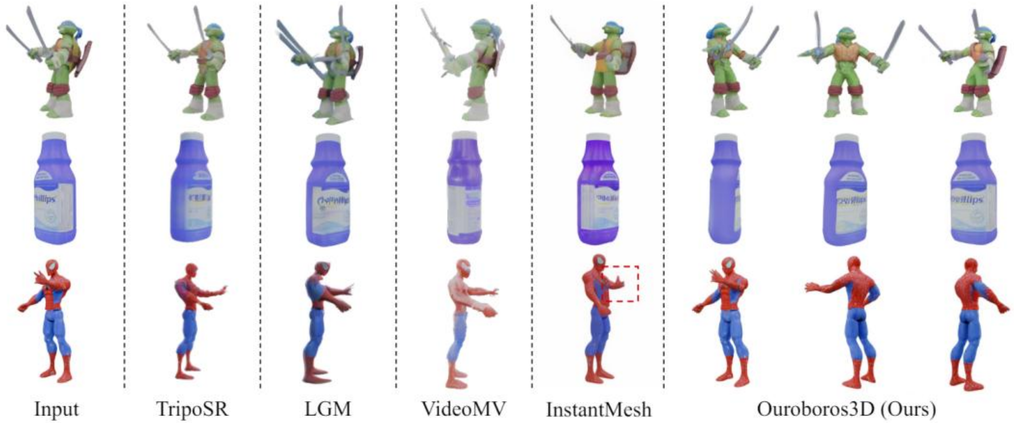
```
begin
    self_cond ← None
    x_t ← Sample from Normal Distribution
    foreach timestep in timesteps do
        pred_x ← F(x_t, cond_image, timestep, cameras, self_cond)
        pred_i ← VAE_Decoder(pred_x)
        self_cond ← G(pred_i, cameras, timestep)
    end
    return pred_i, self_cond
end
```

Comparison Results

Qualitative comparisons of generated multi-view images



Qualitative comparisons for image-to-3D



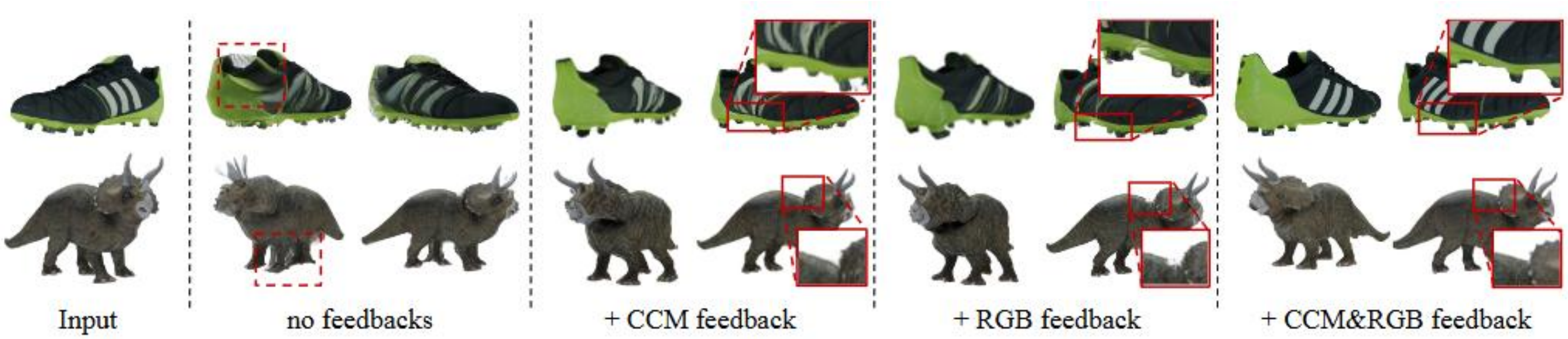
	Method	Resolution	PSNR↑	SSIM↑	LPIPS↓
Image-to-Multiview	SyncDreamer [9]	256 × 256	20.056	0.8163	0.1596
	SV3D [13]	576 × 576	21.042	0.8497	0.1296
	VideoMV [23]	256 × 256	18.605	0.8410	0.1548
	Ouroboros3D (SVD)	512 × 512	**21.770**	**0.8866**	**0.1093**
Image-to-3D	TripoSR [53]	256 × 256	18.481	0.8506	0.1357
	LGM [16]	512 × 512	17.716	0.8319	0.1894
	VideoMV(GS) [23]	256 × 256	18.764	0.8449	0.1569
	InstantMesh (NeRF) [19]	512 × 512	19.948	0.8727	0.1205
	Ouroboros3D (LGM)	512 × 512	**21.761**	**0.8894**	**0.1091**

Table: Quantitative comparison on the quality of generated multi-view images and 3D representation for image-to-multiview and image-to-3D tasks.

GSO Results

Ablation

Qualitative comparison with no-feedback and 3d-aware feedback



Joint Training	CCM Feedback	RGB Feedback	PSNR↑	SSIM↑	LPIPS↓	ΔPSNR↓	ΔSSIM↓	ΔLPIPS↓
✗	✗	✗	20.012	0.8465	0.1287	1.067	0.0125	0.0189
✓	✗	✗	20.549	0.8651	0.1183	0.511	0.0094	0.0070
✓	✓	✗	21.325	0.8937	0.1092	0.304	0.0036	0.0018
✓	✗	✓	21.542	0.8871	0.1103	0.100	0.0101	0.0036
✓	✓	✓	**21.761**	**0.9094**	**0.0991**	**0.009**	**0.0028**	**0.0002**

More Results