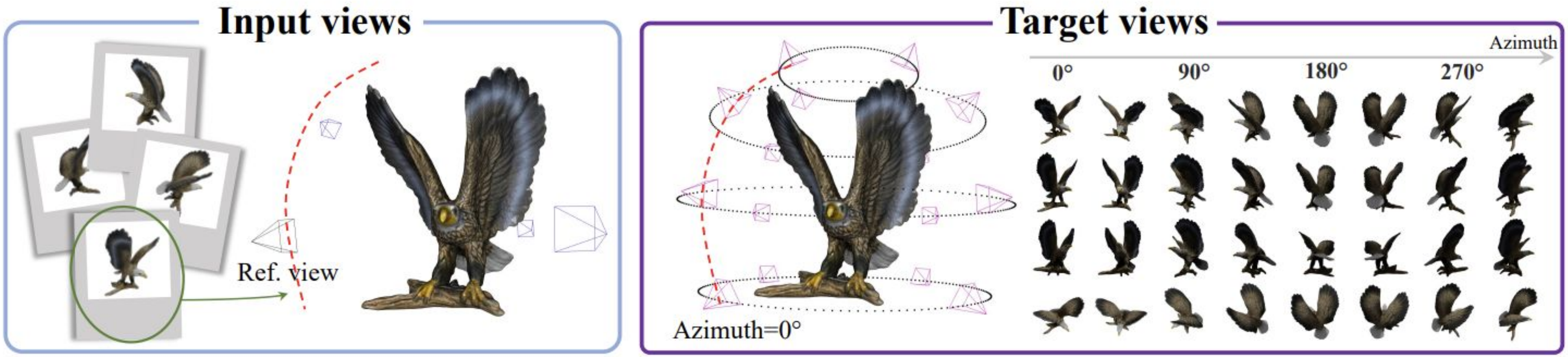
MVDiffusion++: A Dense High-resolution Multi-view Diffusion Model for Single or Sparse-view 3D Object Reconstruction

Shitao Tang, Jiacheng Chen, Dilin Wang, Chengzhou Tang, Fuyang Zhang, Yuchen Fan, Vikas Chandra, Yasutaka Furukawa, Rakesh Ranjan
Simon Fraser University, Meta Reality Labs
ECCV 2024

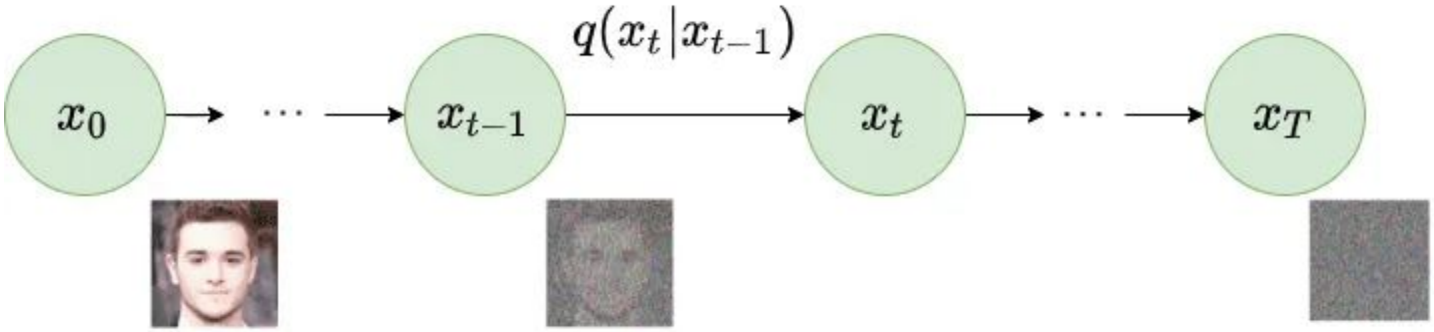
Task: Single or Sparse View Reconstruction

The 32 target images are defined in eight azimuths and four elevation levels.



Forward diffusion process:

- Sample from a basic Gaussian distribution.
- Incremental modifications via Markov chain.
- Structured noise added at each step, controlled by variance schedule.



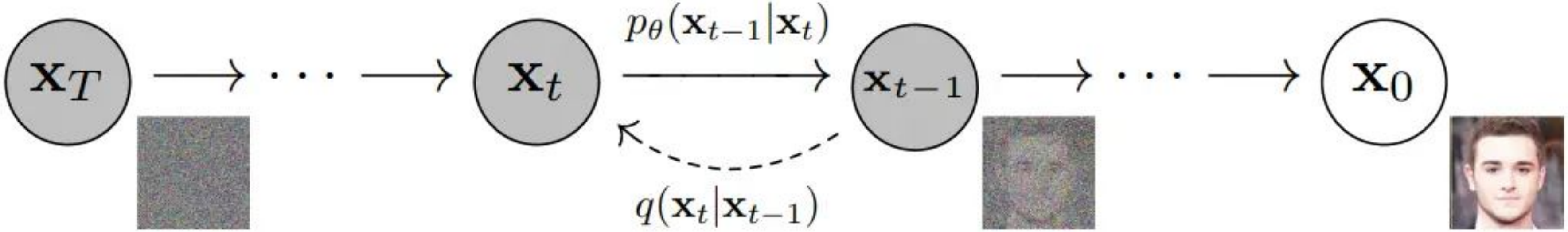
$$q(\mathbf{x}_t|\mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t\mathbf{I}) \quad q(\mathbf{x}_{1:T}|\mathbf{x}_0) = \prod_{t=1}^{T} q(\mathbf{x}_t|\mathbf{x}_{t-1})$$

"Diffusion Models," https://www.superannotate.com/blog/diffusion-models

Diffusion Model

Reverse diffusion process:

- x_T behaves like an isotropic Gaussian distribution.
- Reverse the process to create new data similar to the original dataset.
- Direct calculation of $q(x_{t-1} | x_t)$ is complex.
- Neural network estimates this, adjusting mean and variance.



$$p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) = \mathcal{N}(\mathbf{x}_{t-1}; \boldsymbol{\mu}_\theta(\mathbf{x}_t, t), \boldsymbol{\Sigma}_\theta(\mathbf{x}_t, t))$$

"Diffusion Models," https://www.superannotate.com/blog/diffusion-models

Latent Diffusion Model

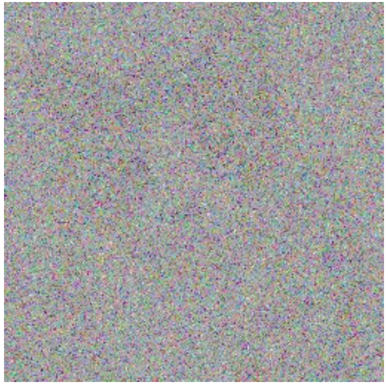
$$L_{LDM} := \mathbb{E}_{\mathcal{E}(\mathbf{x}),\mathbf{y},\epsilon\sim\mathcal{N}(0,1),t}\left[\|\epsilon - \epsilon_\theta(\mathbf{Z}_t, t, \tau_\theta(\mathbf{y}))\|_2^2\right]$$

$\mathbf{Z} = \mathcal{E}(\mathbf{x})$ where $\mathcal{E}$ is the encoder and X is the high-resolution images.

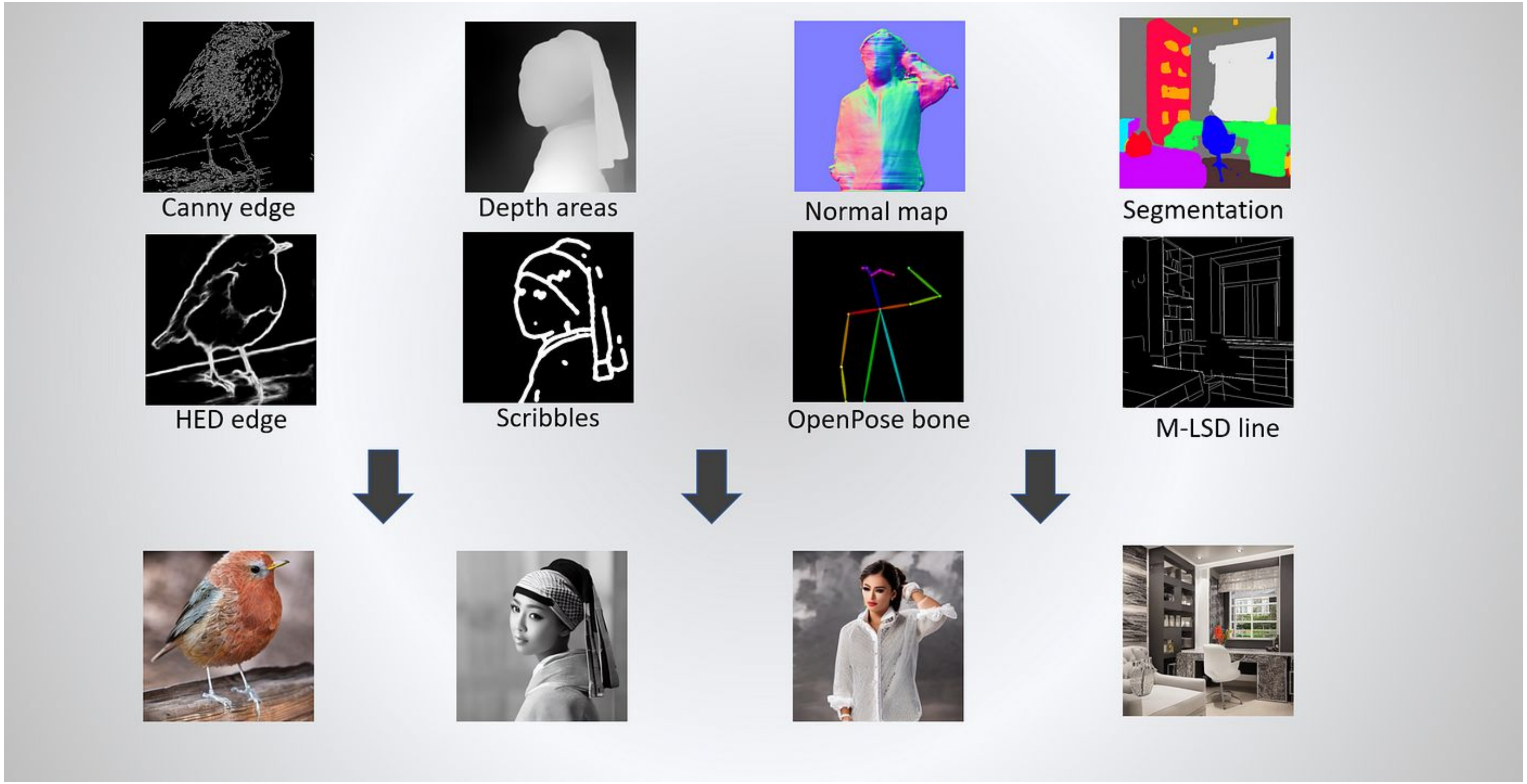
$\mathbf{Z}_t$ is the noisy latent at time step t.

τ_θ is the optional condition encoding.

$\mathbf{y}$ could be a text-prompt, an image, or any other user-specified condition.

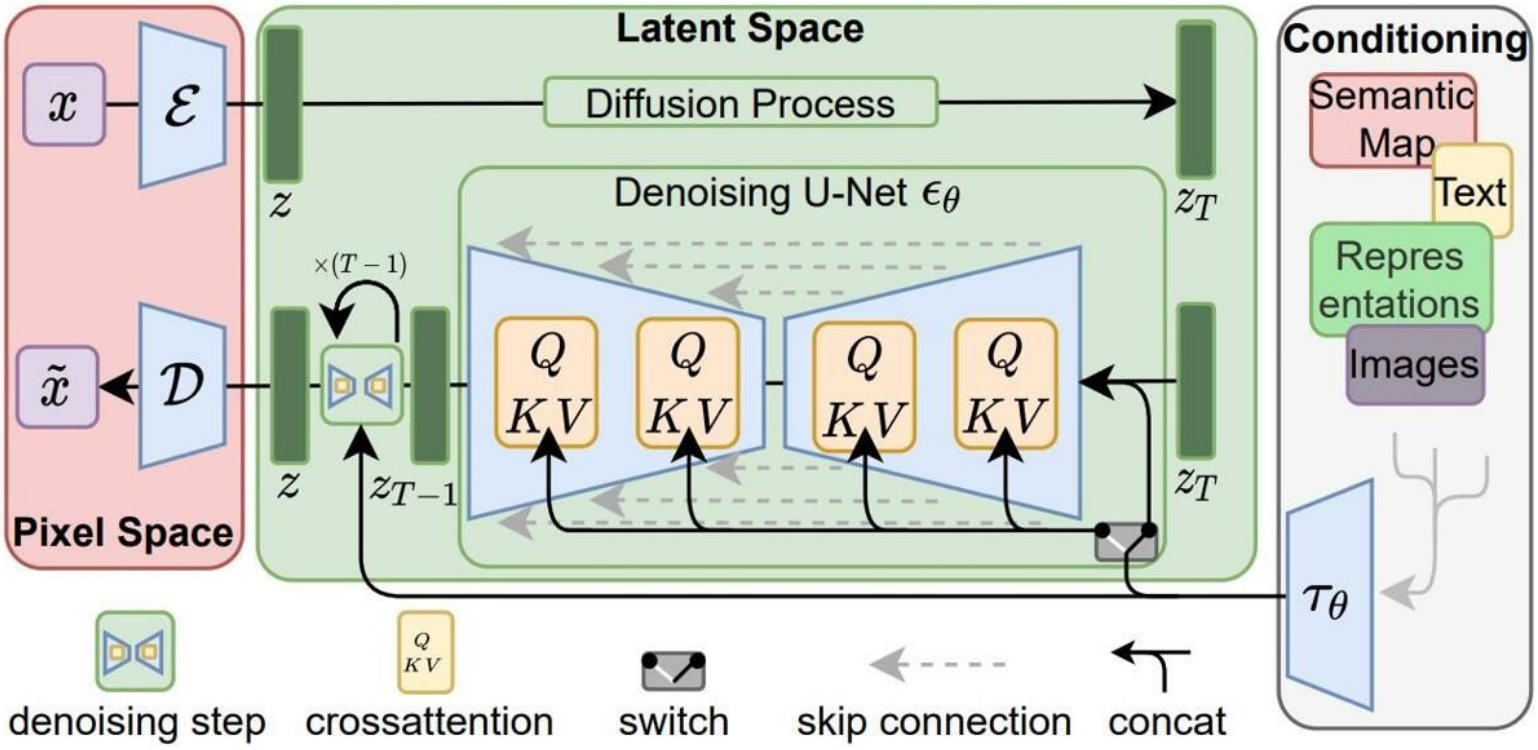


Different Conditions



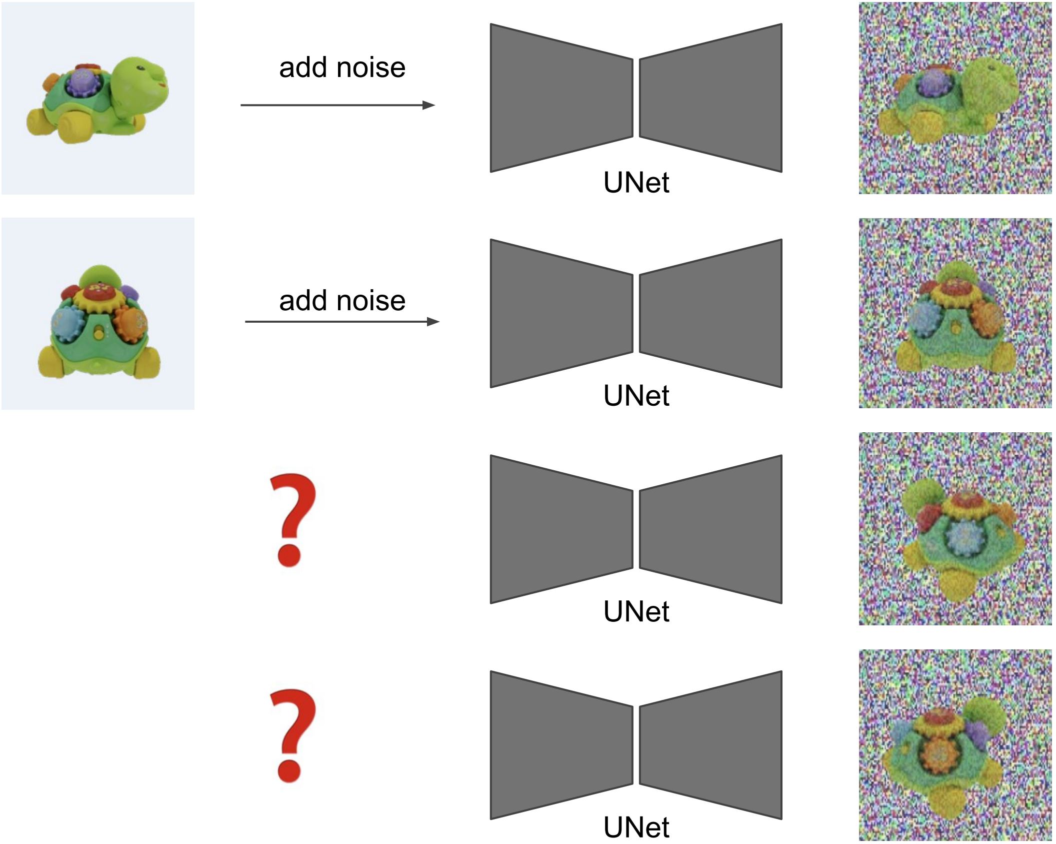
ControlNet (L. Zhang, et al - ICCV 2023)

Stable Diffusion Architecture



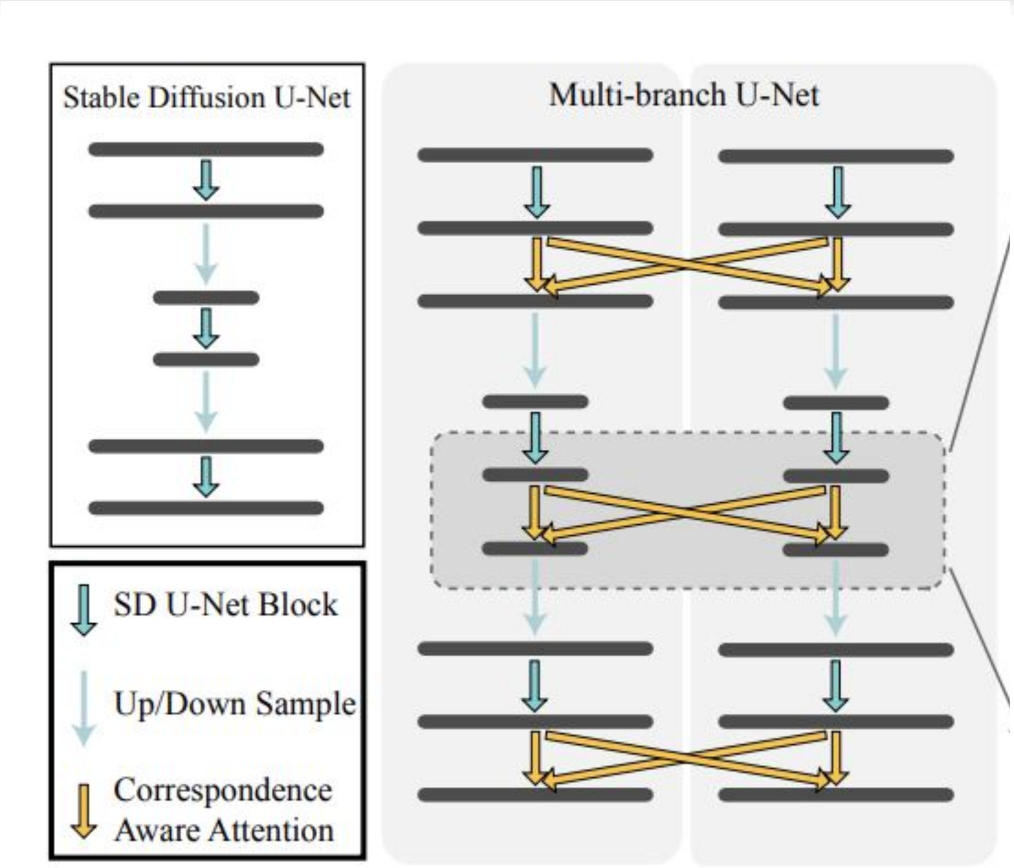
High-Resolution Image Synthesis with Latent Diffusion Models(R. Rombach, et al - CVPR 2022)

Single or Sparse View Reconstruction



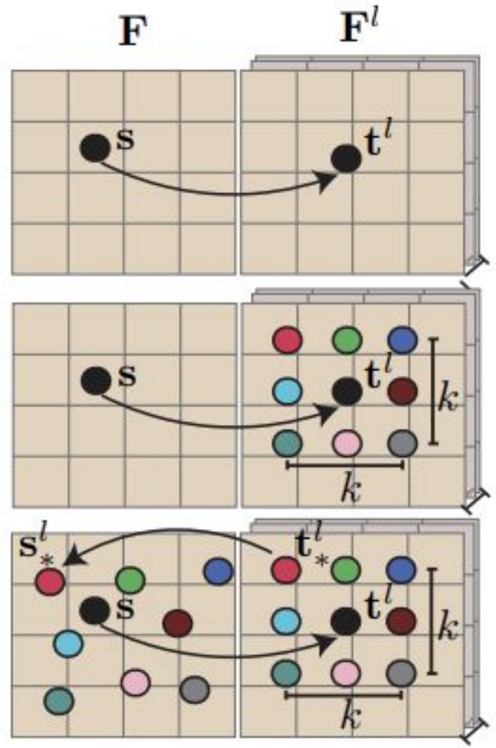
How to connect these?

Correspondence-Aware Attention (CAA)

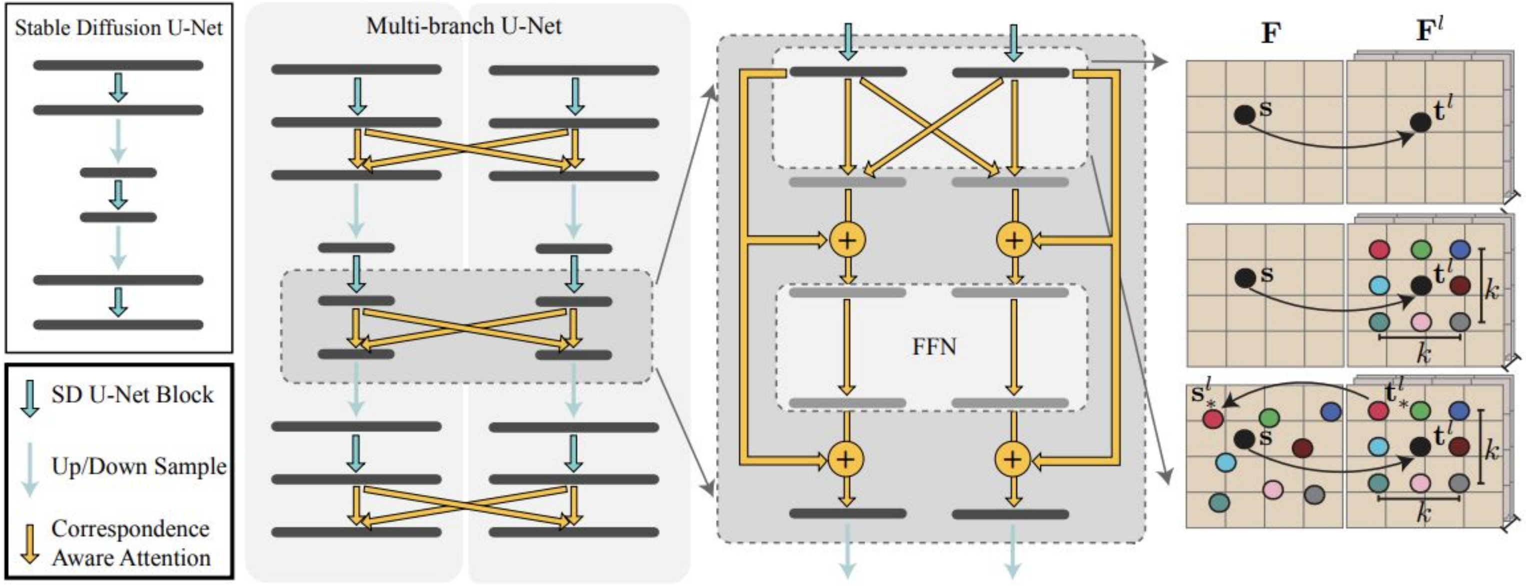


MVDiffusion: Enabling Holistic Multi-view Image Generation with Correspondence-Aware Diffusion(S. Tang, et al - NeurIPS 2023)

Correspondence-Aware Attention (CAA)



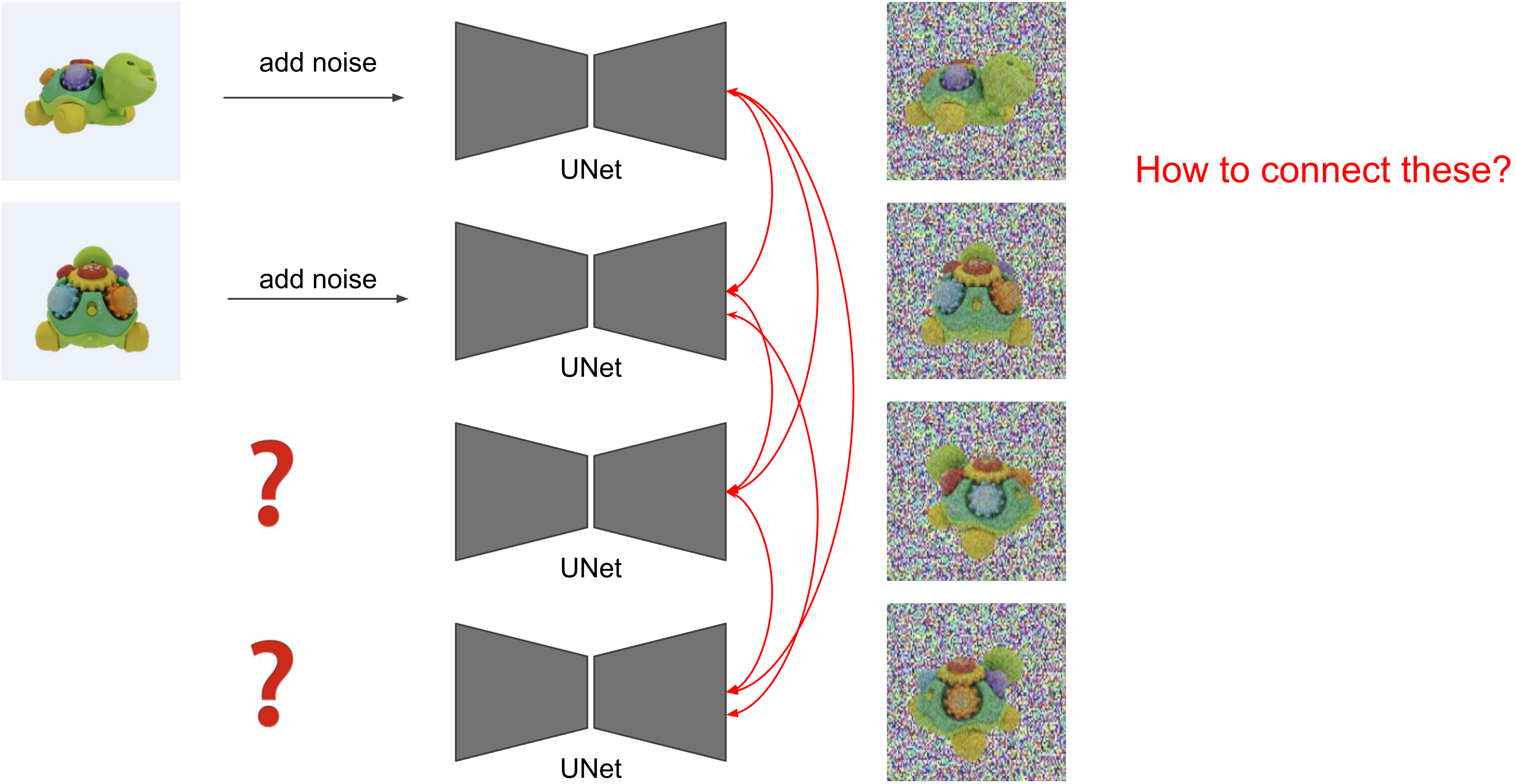
Correspondence-Aware Attention (CAA)



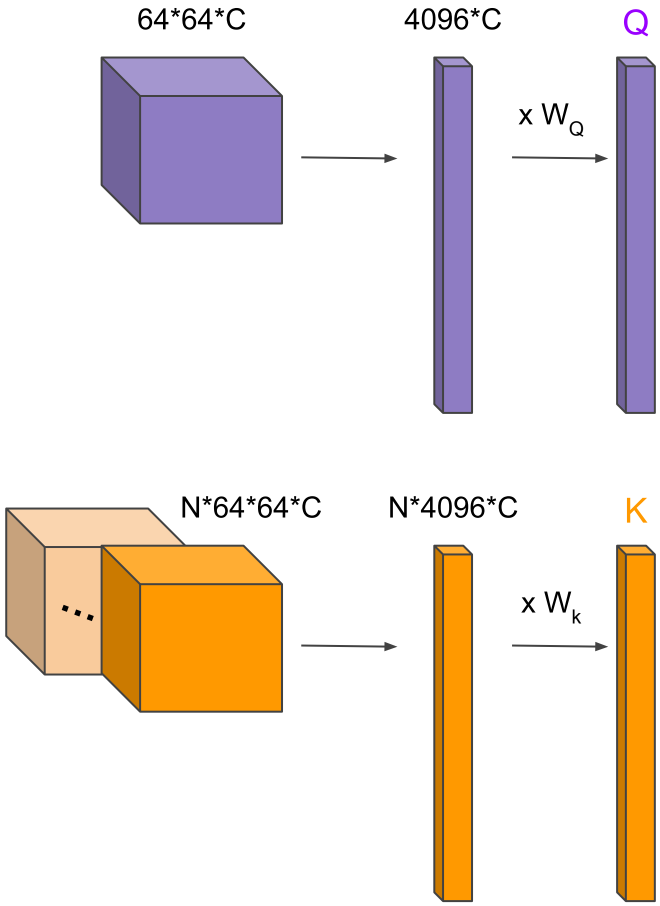
$$\mathbf{M} = \sum \sum \mathrm{SoftMax}\left(\left[\mathbf{W_Q}\bar{\mathbf{F}}(\mathbf{s})\right] \cdot \left[\mathbf{W_K}\bar{\mathbf{F}}^l(t^l_*)\right]\right) \mathbf{W_V}\bar{\mathbf{F}}^l(t^l_*)$$

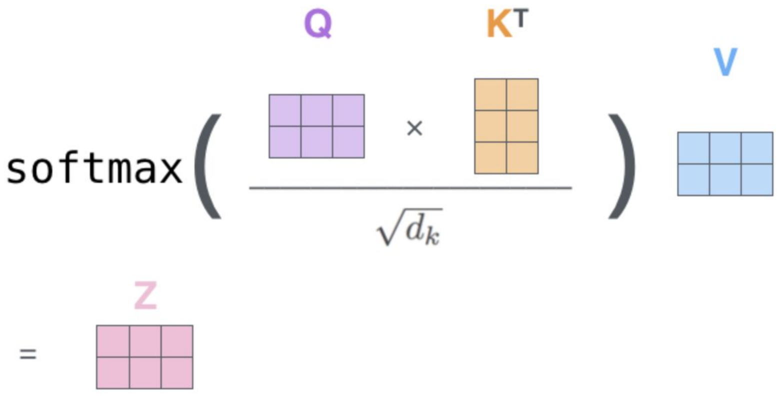
$$\bar{\mathbf{F}}(\mathbf{s}) = \mathbf{F}(\mathbf{s}) + \gamma(0), \quad \bar{\mathbf{F}}^l(t^l_*) = \mathbf{F}^l(t^l_*) + \gamma(\mathbf{s}^l_* - \mathbf{s})$$

Sparse View Reconstruction



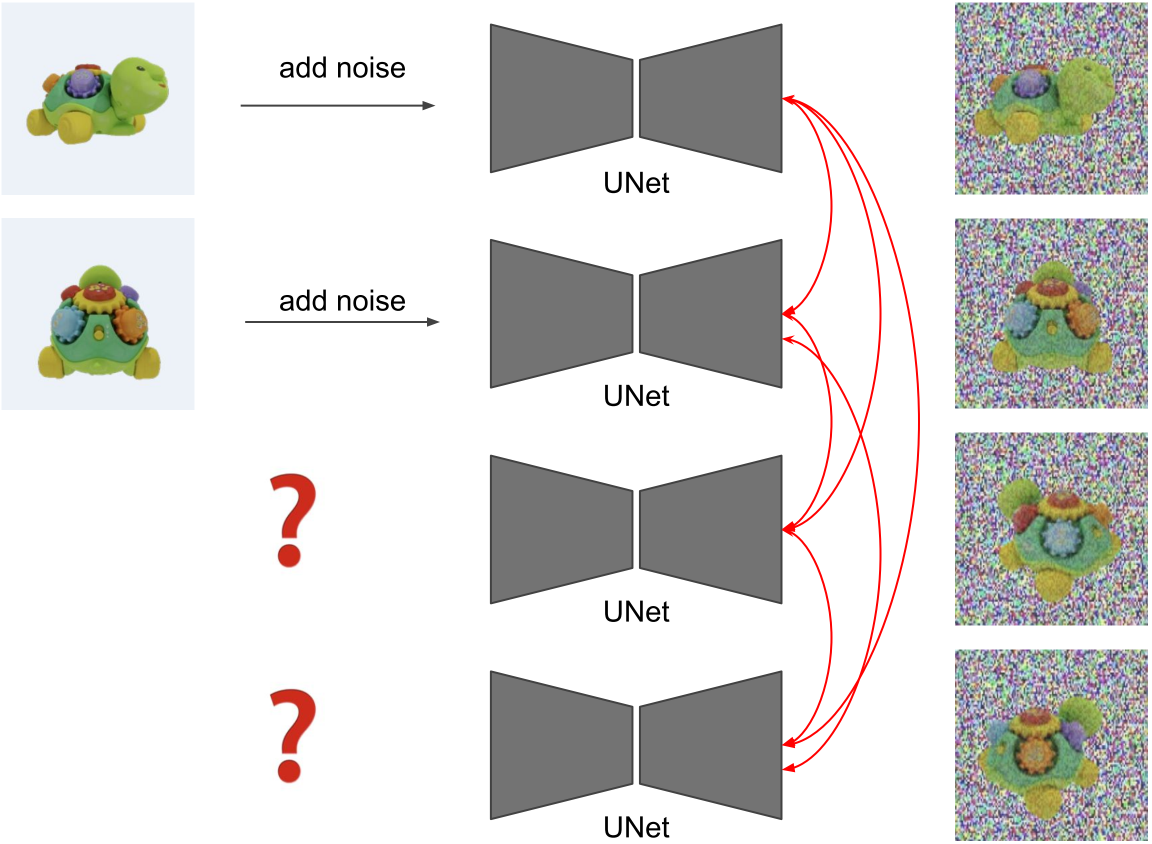
Global Self Attention





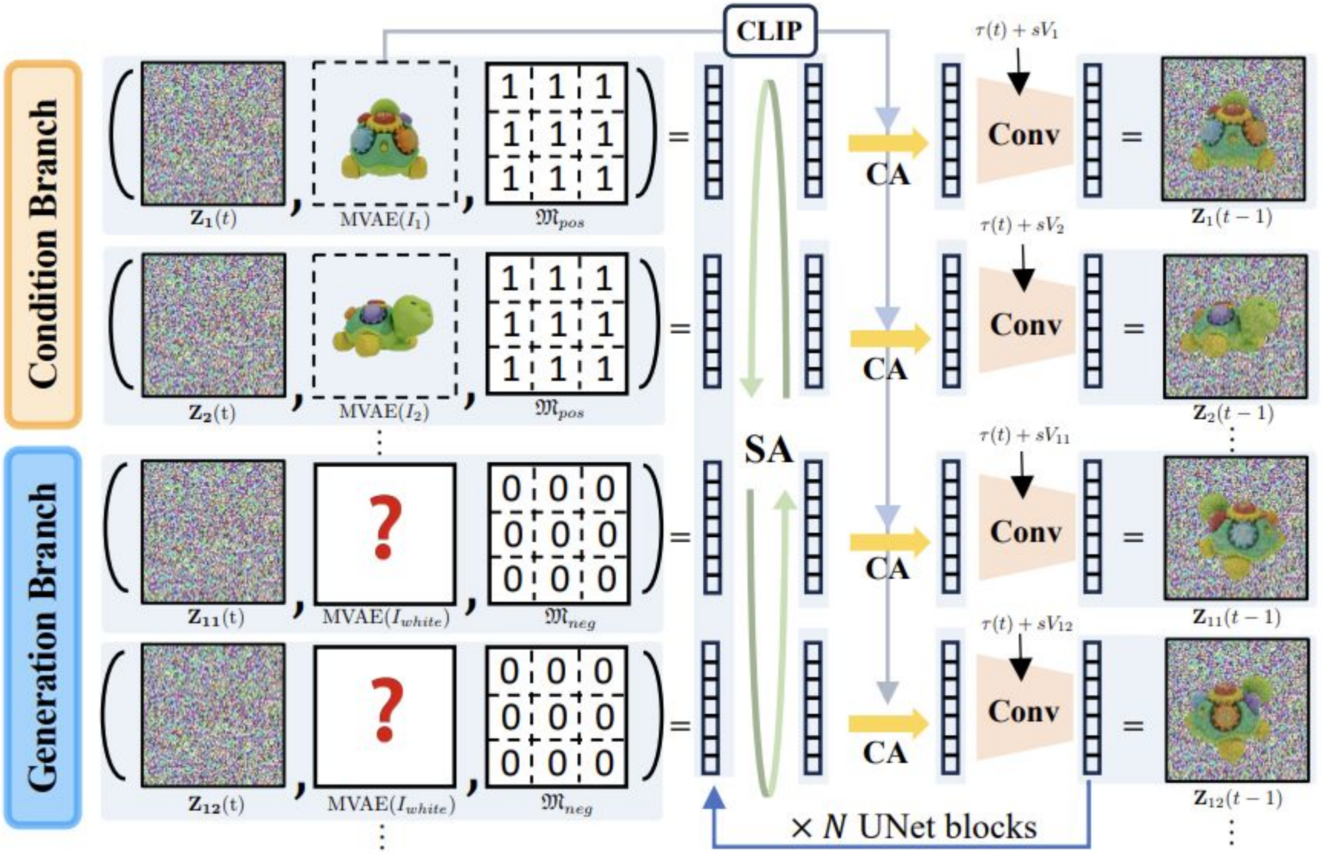
The self-attention calculation in matrix form

Sparse View Reconstruction

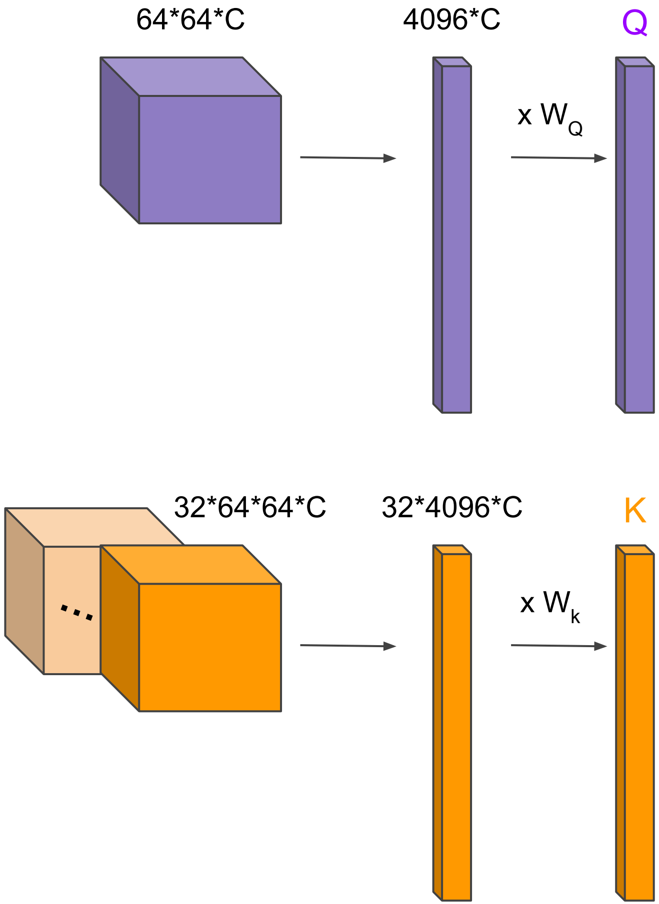


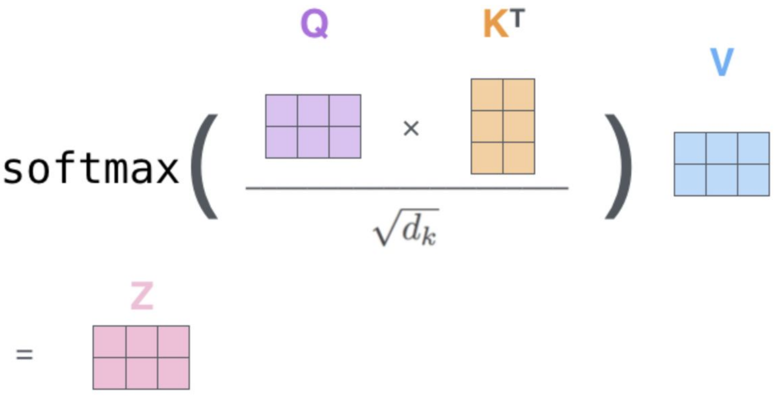
How to connect these?

MVDiffusion++ Architecture



Global Self Attention





The self-attention calculation in matrix form

View dropout training strategy

High Token Count:

- 32 copies of UNet features produce over 130,000 tokens.

Self-Attention Limitation:

- Global self-attention becomes infeasible due to memory constraints, even with advanced memory-efficient transformers.

Solution:

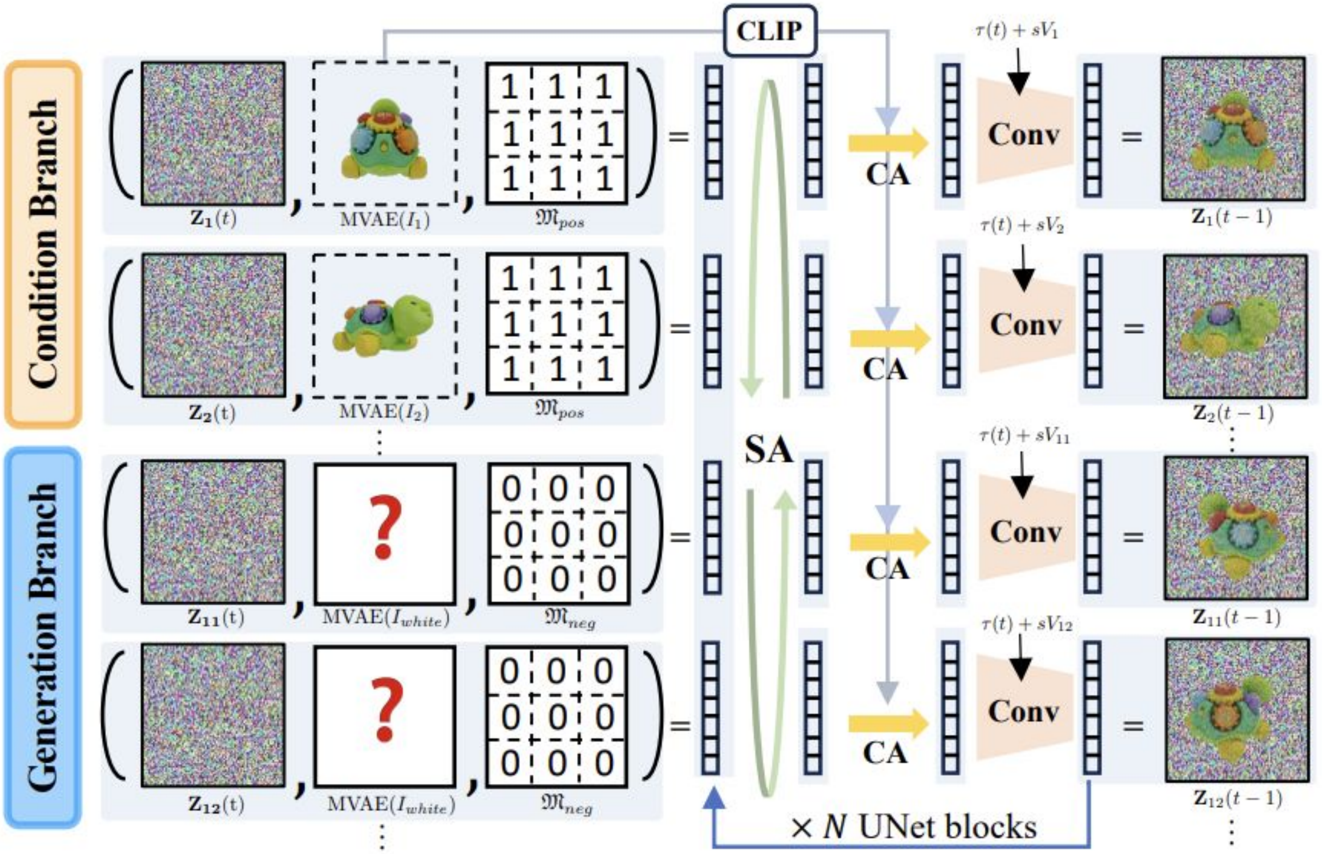
During Training:

- Randomly drop 24 out of 32 views for each object in every iteration.
- Results in significant memory reduction.

At Test Time:

- Utilize full architecture to generate 32 views, ensuring comprehensive output.

MVDiffusion++ Architecture



Key Functions in Each Block

Global Self-Attention:

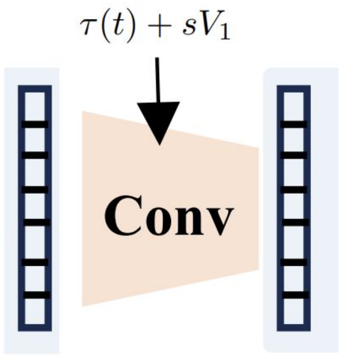
Maintains 3D consistency across all images.

Cross-Attention:

Incorporates CLIP embeddings from the condition images to enhance contextual relevance.

CNN Processing:

Manages per-image features and integrates timestep frequency encoding and image index embeddings.



$Z(t)$: Noisy latent
U : Feature map
I : Image
M : Background/Foreground mask
$\mathfrak{M}_{neg}$: Zero-mask
$\mathfrak{M}_{pos}$: One-mask
CAA : CAA attention
SA/CA : Self/Cross attention
CLIP : CLIP encoder
CNN : Convolution network
MVAE : Mask-aware VAE

MVDiffusion Block

[At first block]

$\forall i\; U_i^0 \leftarrow \mathbf{CNN}(Z_i(t))$

[For each block]

$\forall i\; U_i^b \leftarrow \mathbf{CAA}(U_i^b, \{U_{i-1}^b, U_{i+1}^b\})$

$\forall i\; U_i^b \leftarrow \mathbf{SA}(U_i^b)$

$\forall i\; U_i^b \leftarrow \mathbf{CA}(U_i^b, \mathbf{CLIP}(T_{text}))$

$\forall i\; U_i^{b+1} \leftarrow \mathbf{CNN}([U_i^b, \tau(t)+s\,\tau(V_i)])$

[At last block]

$\forall i\; Z_i(t-1) \leftarrow \mathrm{DDPM}(\mathbf{CNN}(U_i^{bmax}))$

MVDiffHD Block

[At first block]

$$\forall i\; U_i^0 \leftarrow \begin{cases} \mathbf{CNN}([Z_i(t), \mathbf{MVAE}(I_i, M_i), \mathfrak{M}_{pos}]), \text{ conditional branch,} \\ \mathbf{CNN}([Z_i(t), \mathbf{MVAE}(I_{white}, \mathfrak{M}_{neg}), \mathfrak{M}_{neg}]), \text{ generation branch.} \end{cases}$$

[For each block]

$\forall i\; U_i^b \leftarrow \mathbf{SA}(\{U_*^b\})$

$\forall i\; U_i^b \leftarrow \mathbf{CA}(U_i^b, \mathbf{CLIP}(I_i \in I_{cond}))$

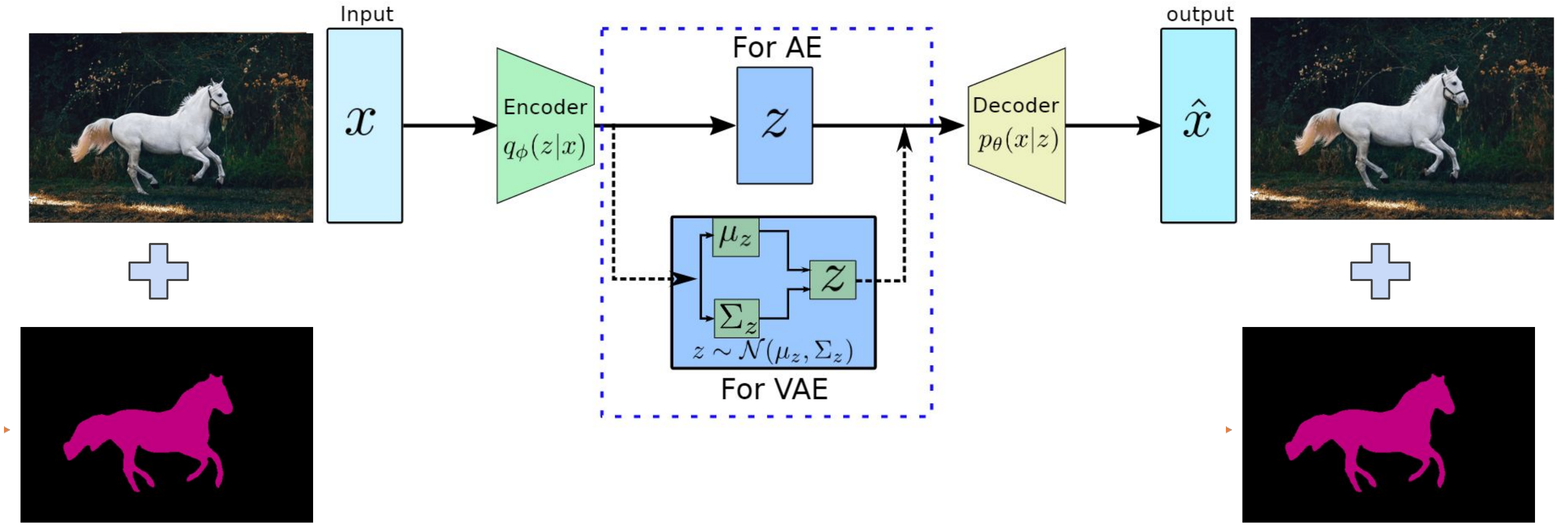
$\forall i\; U_i^{b+1} \leftarrow \mathbf{CNN}([U_i^b, \tau(t)+s\,\tau(V_i)])$

[At last block]

$\forall i\; Z_i(t-1) \leftarrow \mathrm{DDPM}(\mathbf{CNN}(U_i^{bmax}))$

Mask-aware VAE pre-fine-tuning

- Adapts VAE to process object images with segmentation masks.
- Dataset Used: Approximately 3 million RGBA images from the Objaverse dataset.
- Enhances PSNR from 36.6 to 41.2, improving quality of generated 3D models.



Training Resource

Hardware Used:

- Utilized 128 Nvidia H100 GPUs.

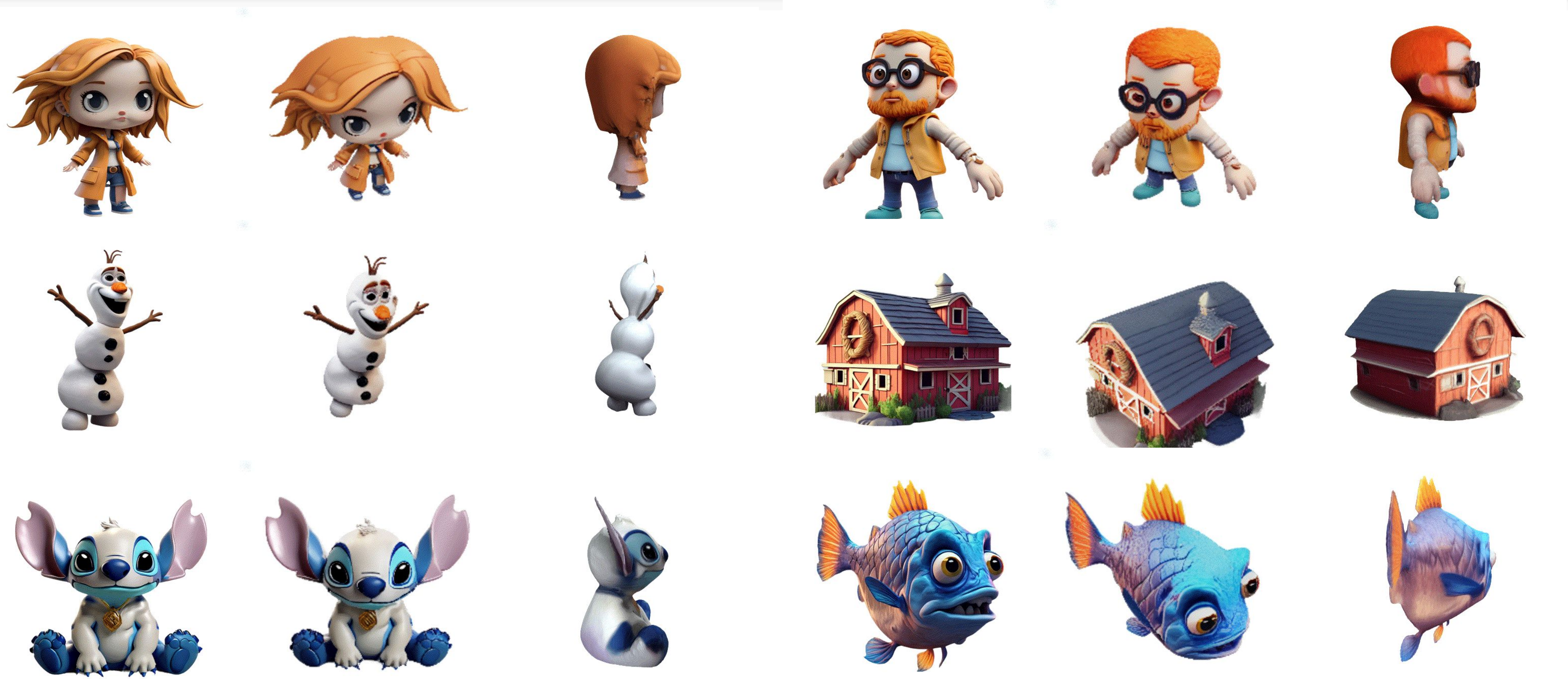
Duration of Training:

- Training conducted continuously for approximately one week.

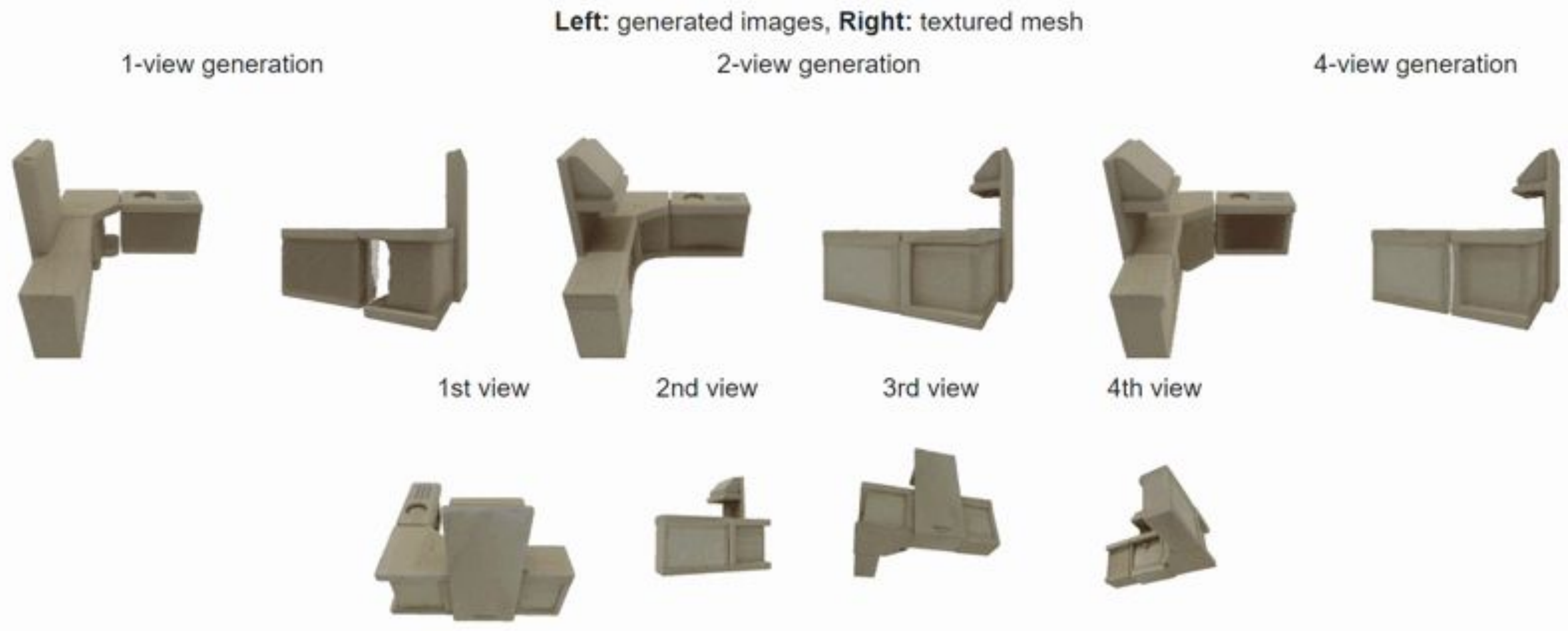
Results

Task →	3D reconstruction		Novel view synthesis		
Method	Chamfer Dist.↓	Vol. IoU↑	PSNR↑	SSIM↑	LPIPS↓
Realfusion [20]	0.0819	0.2741	15.26	0.722	0.283
Magic123 [26]	0.0516	0.4528	-	-	-
One-2-3-45 [16]	0.0629	0.4086	-	-	-
Point-E [24]	0.0426	0.2875	-	-	-
Shap-E [13]	0.0436	0.3584	-	-	-
Zero123 [17]	0.0339	0.5035	18.93	0.779	0.166
SyncDreamer [18]	0.0261	0.5421	20.05	0.798	0.146
Wonder3D [19]*	0.0329	0.5768	-	-	-
Open-LRM [9]*	0.0285	0.5945	-	-	-
Ours	**0.0165**	**0.6973**	**21.45**	**0.844**	**0.129**

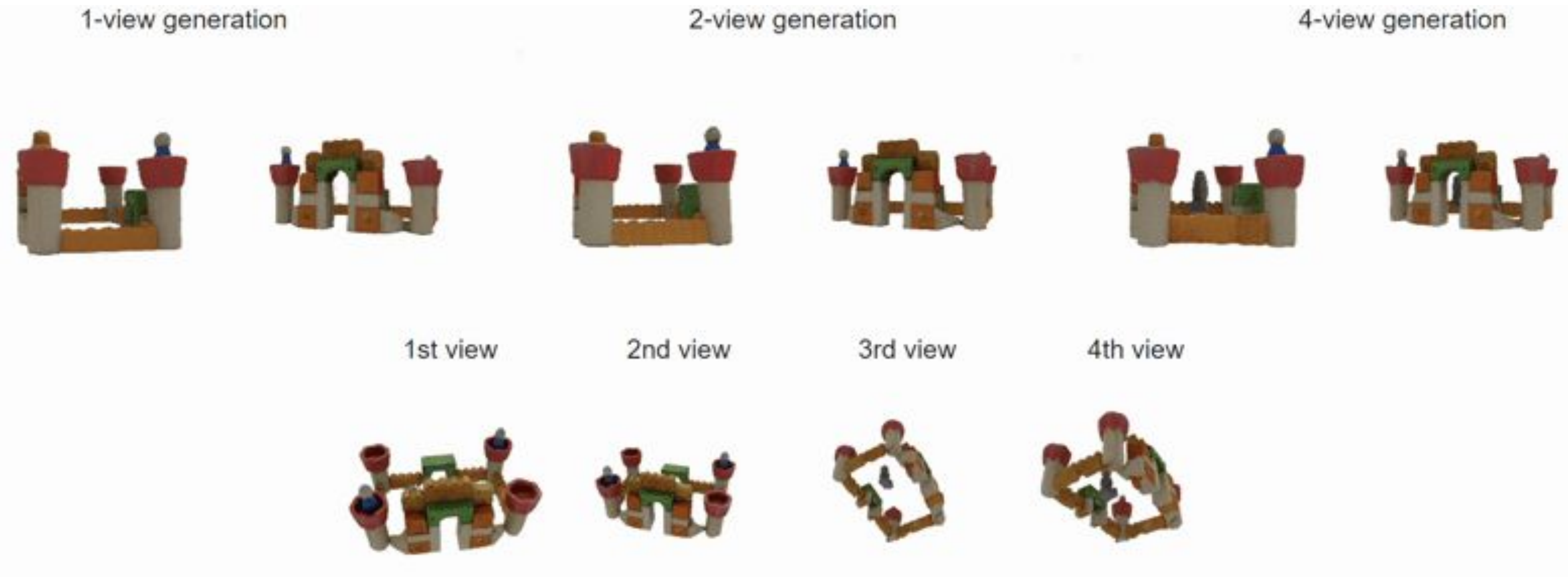
Single view reconstruction



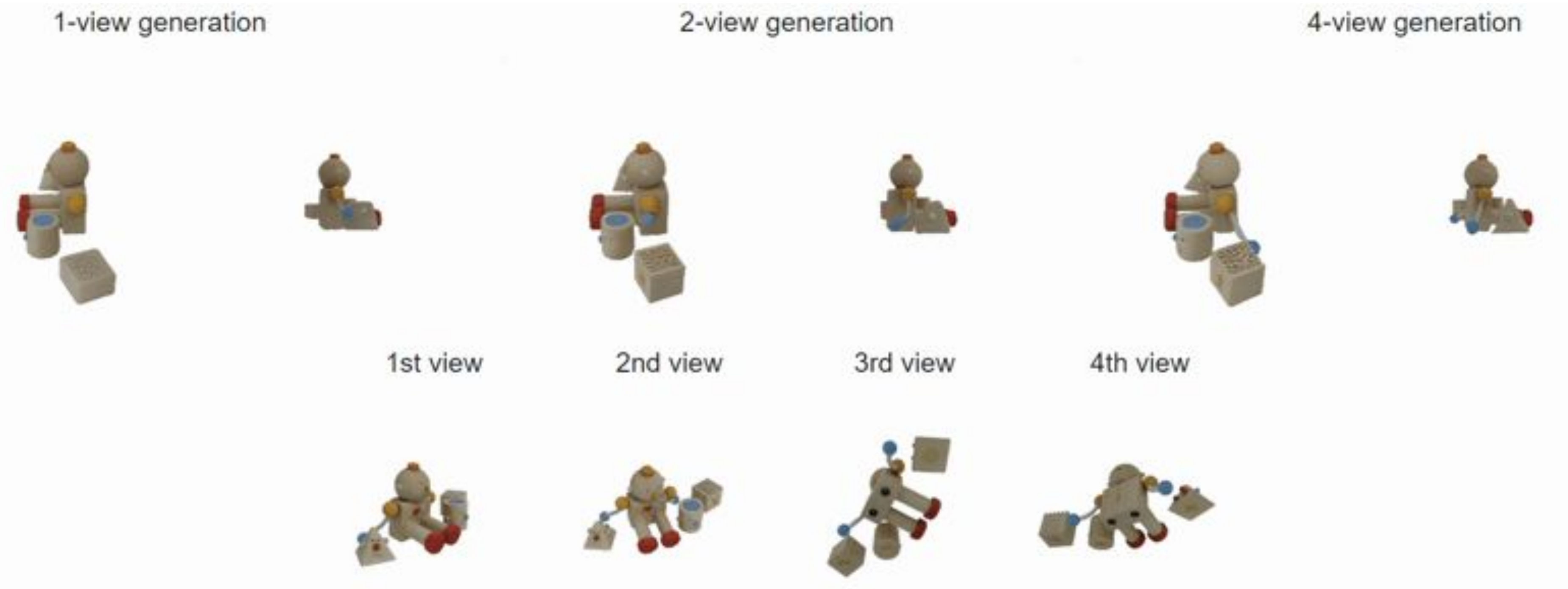
Sparse view reconstruction



Sparse view reconstruction



Sparse view reconstruction



Thank you!

- Thank you for your attention!
- I appreciate your time and interest.
- If you have any questions, please feel free to ask.
- Contact information: alimohammadiamirhossein@gmail.com