LaRender: Training-Free Occlusion Control in Image Generation via Latent Rendering

ICCV 2025 (***Oral & Award Candidate***)

Xiaohang Zhan

Dingming Liu

Tencent

Speaker: Xiaohang Zhan
(Tencent => Adobe)

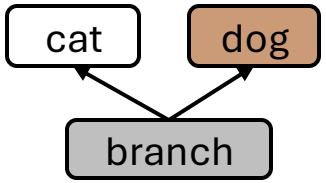
Text-to-image models struggle in occlusion control

SDXL

FLUX

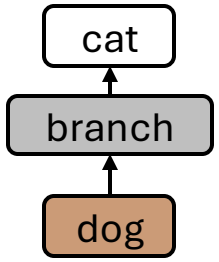
Nano Banana

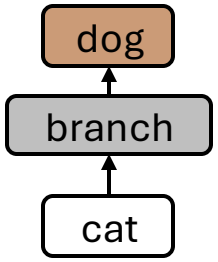
Prompt: In a forest, a white cat and a brown dog are standing, a long branch is in front of them and occludes both of them.



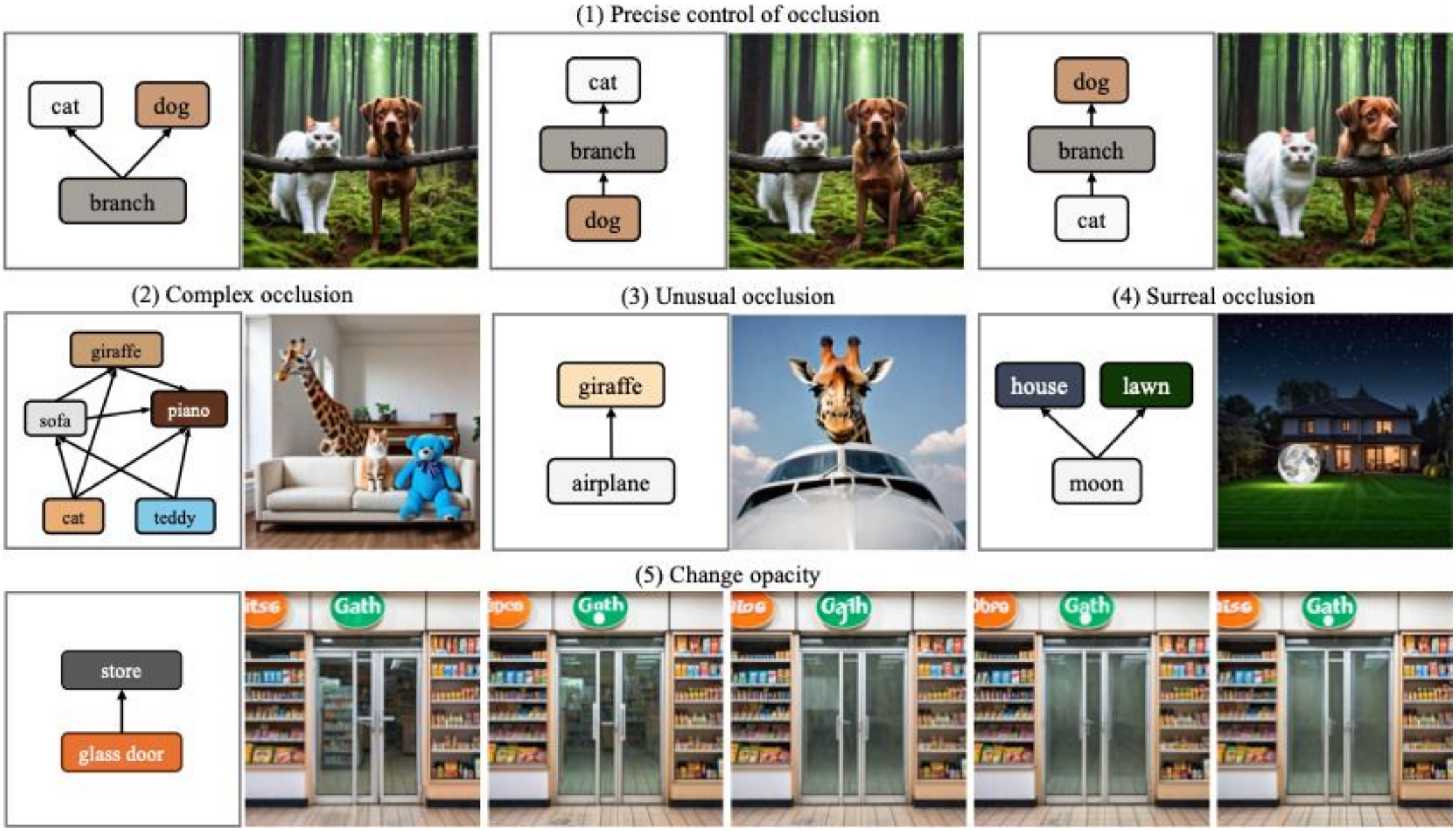
Text-to-image models struggle in occlusion control

Nano Banana

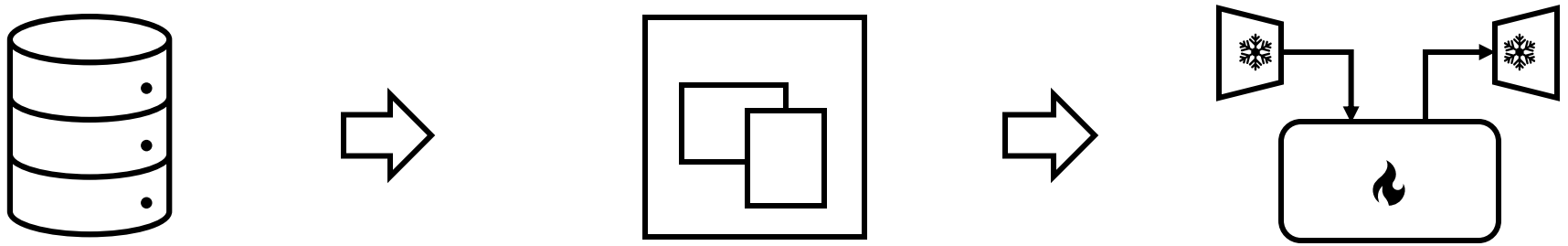




What does LaRender do?



A popular way



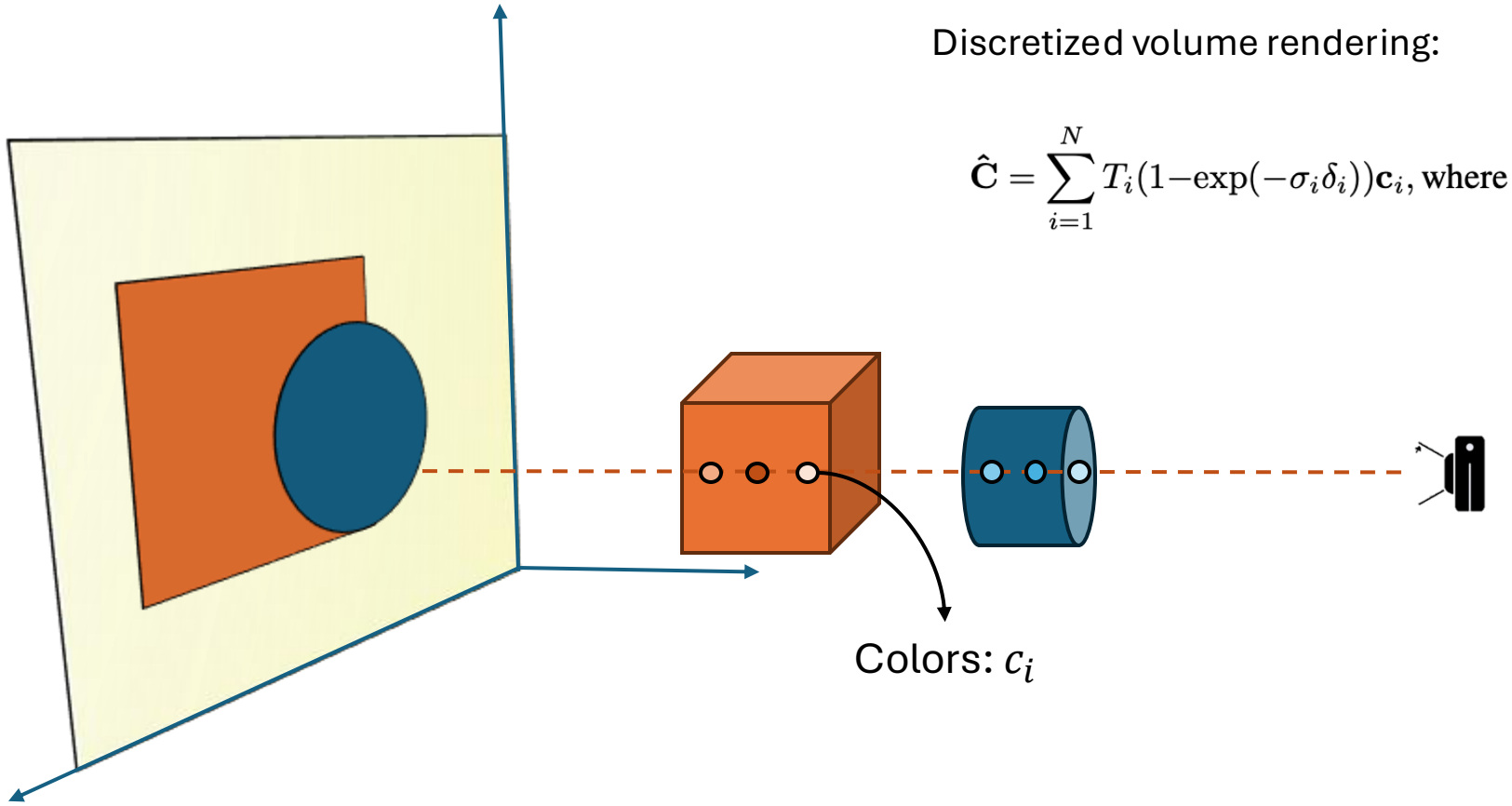
Data curation

Control signal design

Model finetuning

It would work, but the cost is high.

What is visual occlusion?



Discretized volume rendering:

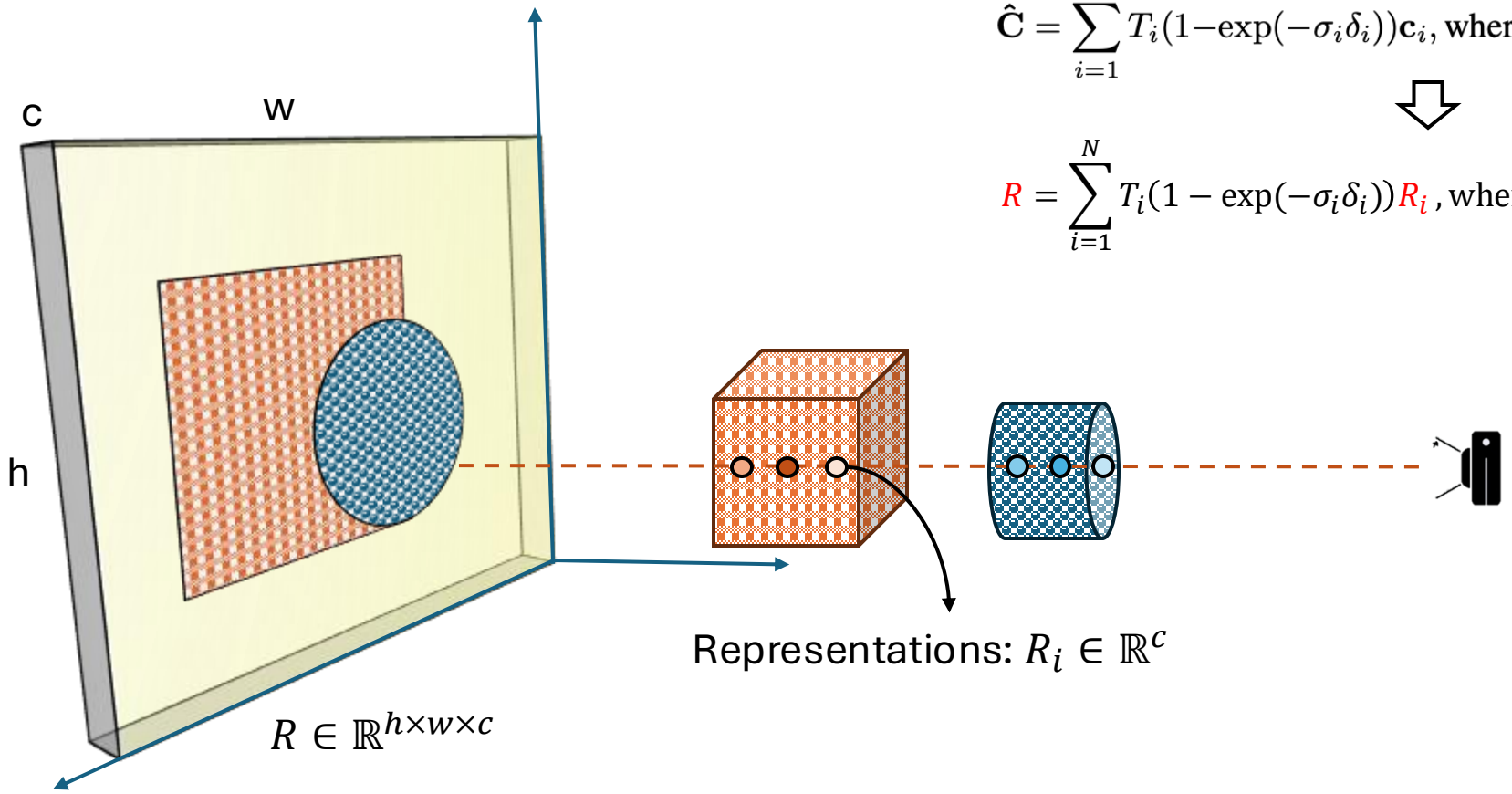
$$\hat{\mathbf{C}} = \sum_{i=1}^{N} T_i(1-\exp(-\sigma_i\delta_i))\mathbf{c}_i, \text{where } T_i = \exp(-\sum_{j=1}^{i-1}\sigma_j\delta_j),$$

🤔 What if they are high dimensional representations?

From physical rendering to latent rendering

$$\hat{\mathbf{C}} = \sum_{i=1}^{N} T_i(1-\exp(-\sigma_i\delta_i))\mathbf{c}_i, \text{ where } T_i = \exp(-\sum_{j=1}^{i-1}\sigma_j\delta_j),$$

$$R = \sum_{i=1}^{N} T_i(1-\exp(-\sigma_i\delta_i))R_i, \text{ where } T_i = \exp(-\sum_{j=1}^{i-1}\sigma_j\delta_j)$$



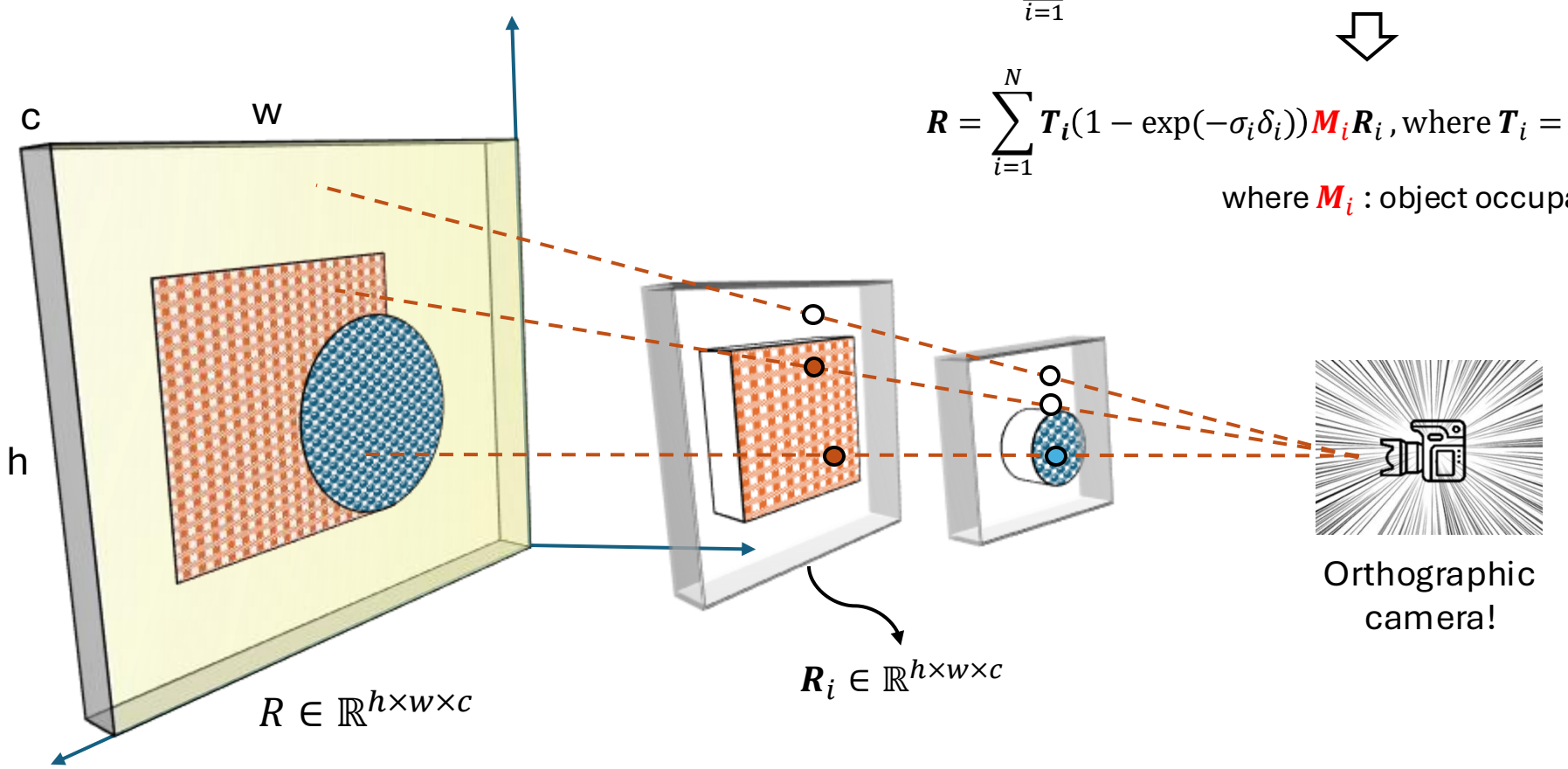
3D representations are more difficult to obtain than 2D ones. Let's simplify them.

From physical rendering to latent rendering

$$R = \sum_{i=1}^{N} T_i(1 - \exp(-\sigma_i \delta_i))R_i \text{, where } T_i = \exp(-\sum_{j=1}^{i-1} \sigma_j \delta_j)$$

$$\boldsymbol{R} = \sum_{i=1}^{N} \boldsymbol{T_i}(1 - \exp(-\sigma_i \delta_i))\boldsymbol{M_i}\boldsymbol{R_i} \text{, where } \boldsymbol{T_i} = \exp(-\sum_{j=1}^{i-1} \boldsymbol{M_j}\sigma_j \delta_j)$$

where $\boldsymbol{M_i}$: object occupancy mask



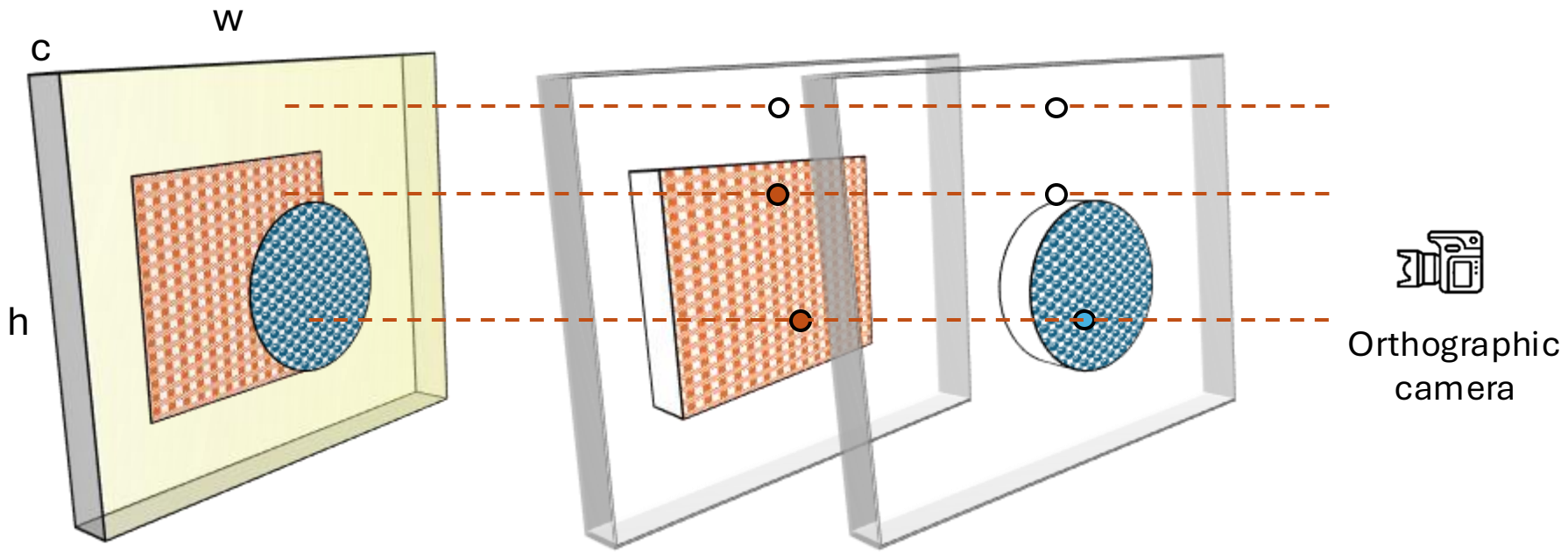
Using 2D hidden states from text-to-image diffusion models while keeping 3D layouts and the shape of the cross sections.

🤔 But how to determine their distances to the camera, and the FOV of the camera?

From physical rendering to latent rendering

$$\boldsymbol{R} = \sum_{i=1}^{N} \boldsymbol{T}_i (1 - \exp(-\sigma_i \delta_i)) \boldsymbol{M}_i \boldsymbol{R}_i \,, \text{where } \boldsymbol{T}_i = \exp(-\sum_{j=1}^{i-1} \boldsymbol{M}_j \sigma_j \delta_j)$$

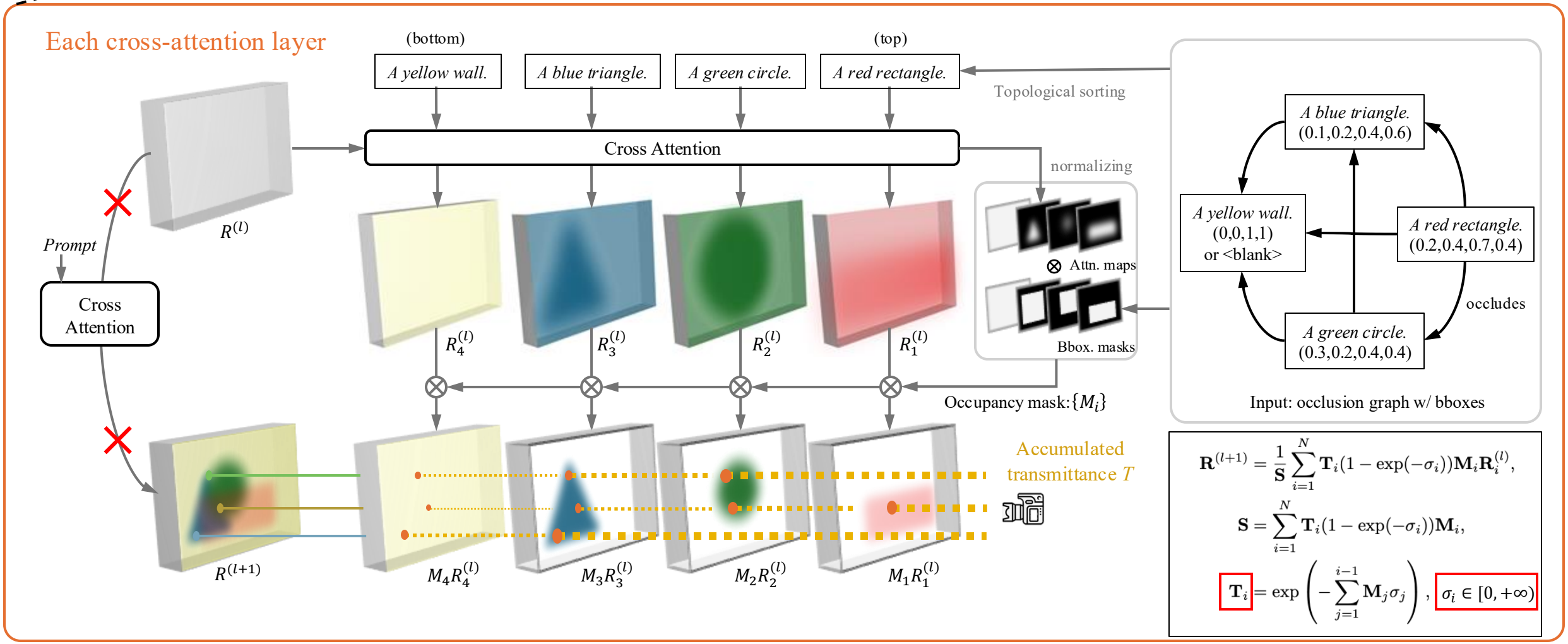
$$\boldsymbol{R} = \sum_{i=1}^{N} \boldsymbol{T}_i (1 - \exp(-\sigma_i)) \boldsymbol{M}_i \boldsymbol{R}_i \,, \text{where } \boldsymbol{T}_i = \exp(-\sum_{j=1}^{i-1} \boldsymbol{M}_j \sigma_j)$$



Now we created the foundation of Latent Rendering!

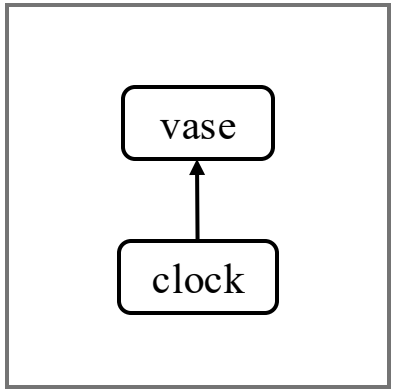
LaRender: a non-parametric framework

Pretrained text-to-image diffusion model

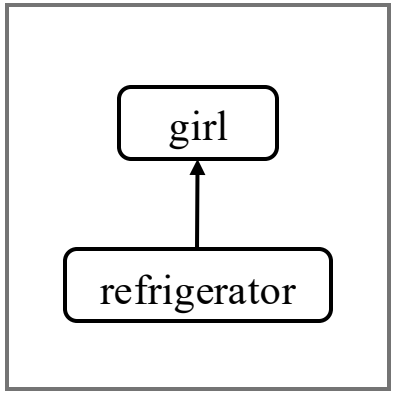


LaRender results

Prompt: A vase hidden by a clock.



Prompt: A girl hidden by a refrigerator.

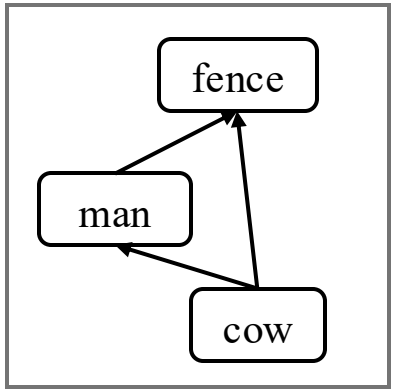


SDXL (text control)	FLUX.1-dev (text control)	MIGC (layout control)	3DIS (layout control)	LaRender (occlusion control)

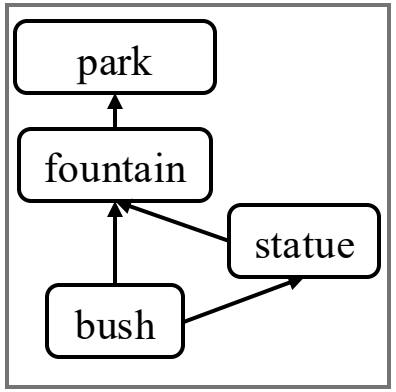
Comparison

LaRender results

Prompt: A cow occludes a man and a fence, the man occludes the fence.



Prompt: In a park, a fountain is partially obscured by a lion statue, and in front of them is a bush that hides both of them.



SDXL (text control) | FLUX.1-dev (text control) | MIGC (layout control) | 3DIS (layout control) | LaRender (occlusion control)

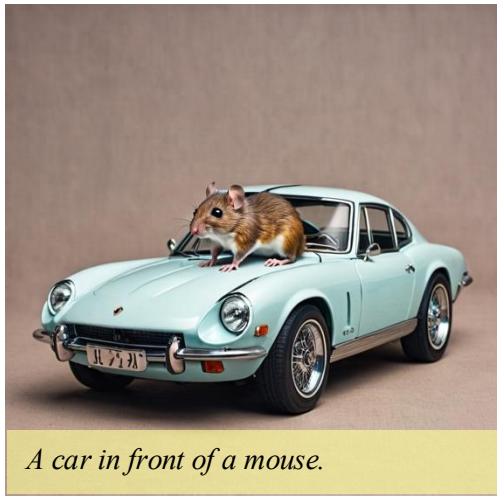
Comparison

LaRender results

Similar layout, different occlusion

LaRender results

Inaccurate position



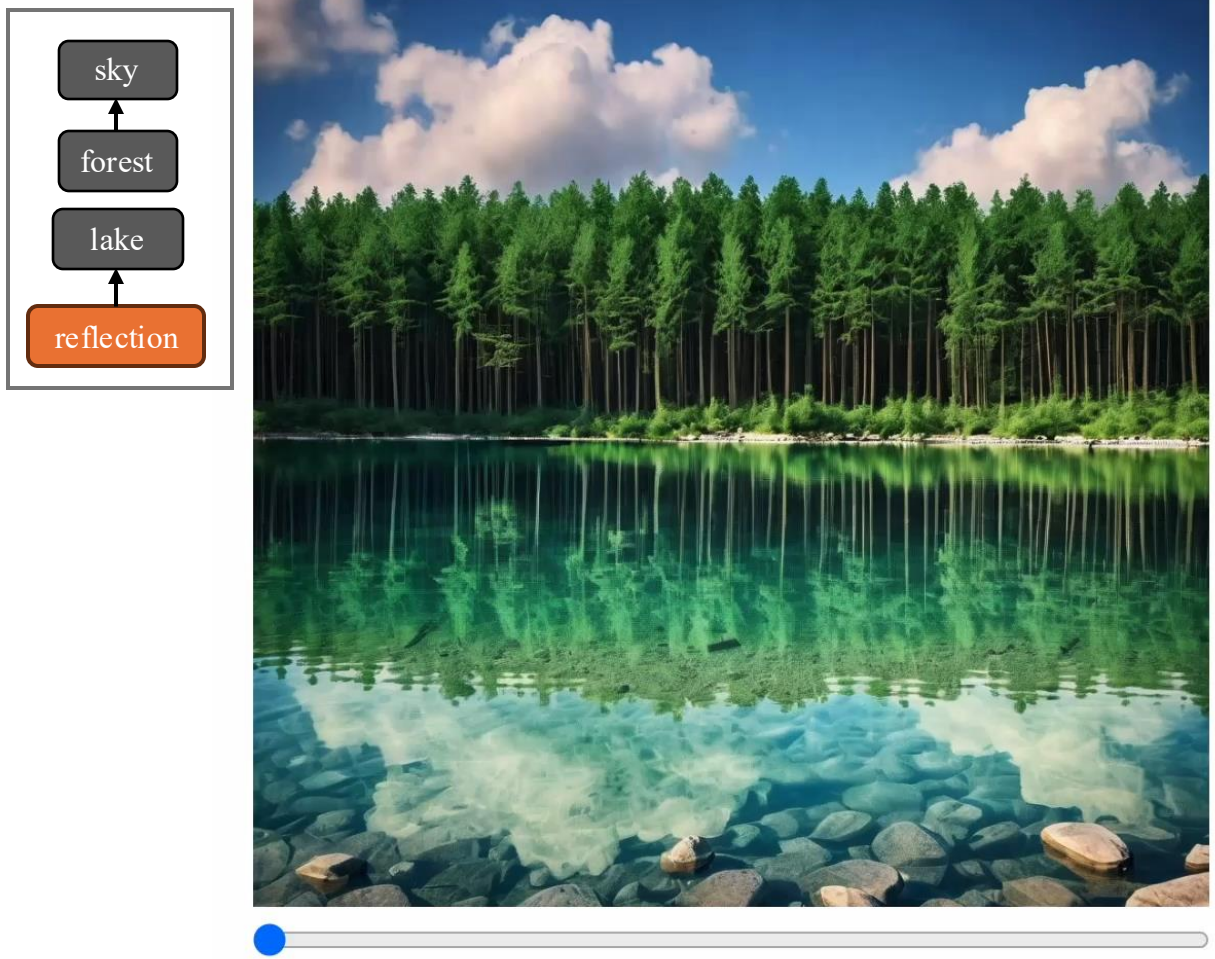
Wrong occlusion

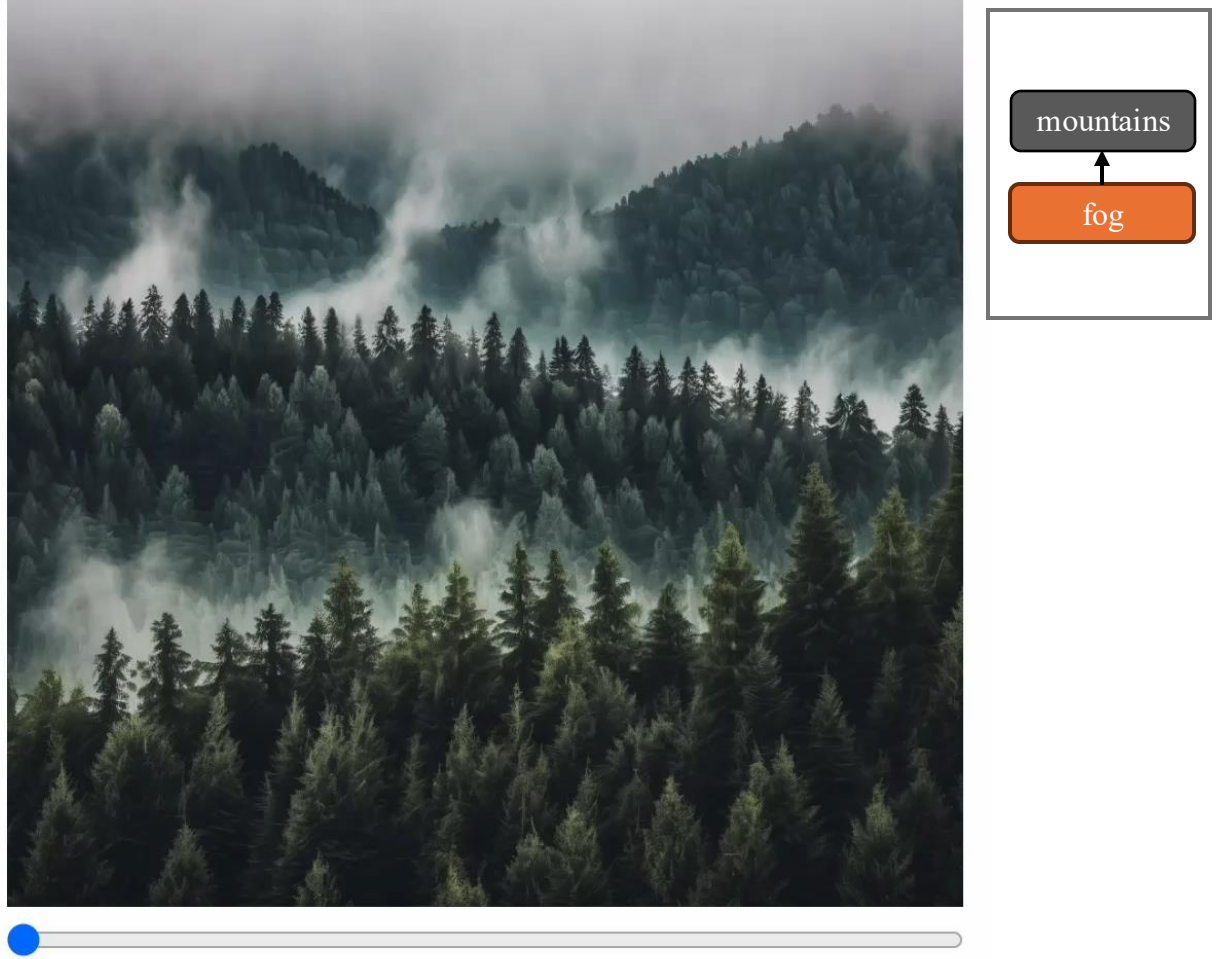
Concept lost

Concept mixed

Failure cases

Semantic opacity control





$$\alpha_i = 1 - \exp(-\sigma_i) \in [0, 1)$$

Semantic opacity control

cathedral

palm trees

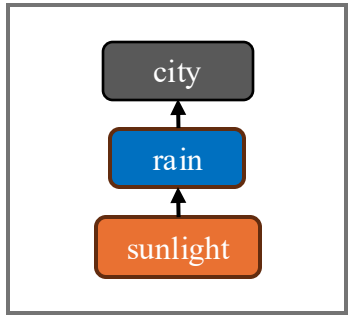
palm trees: 0.00

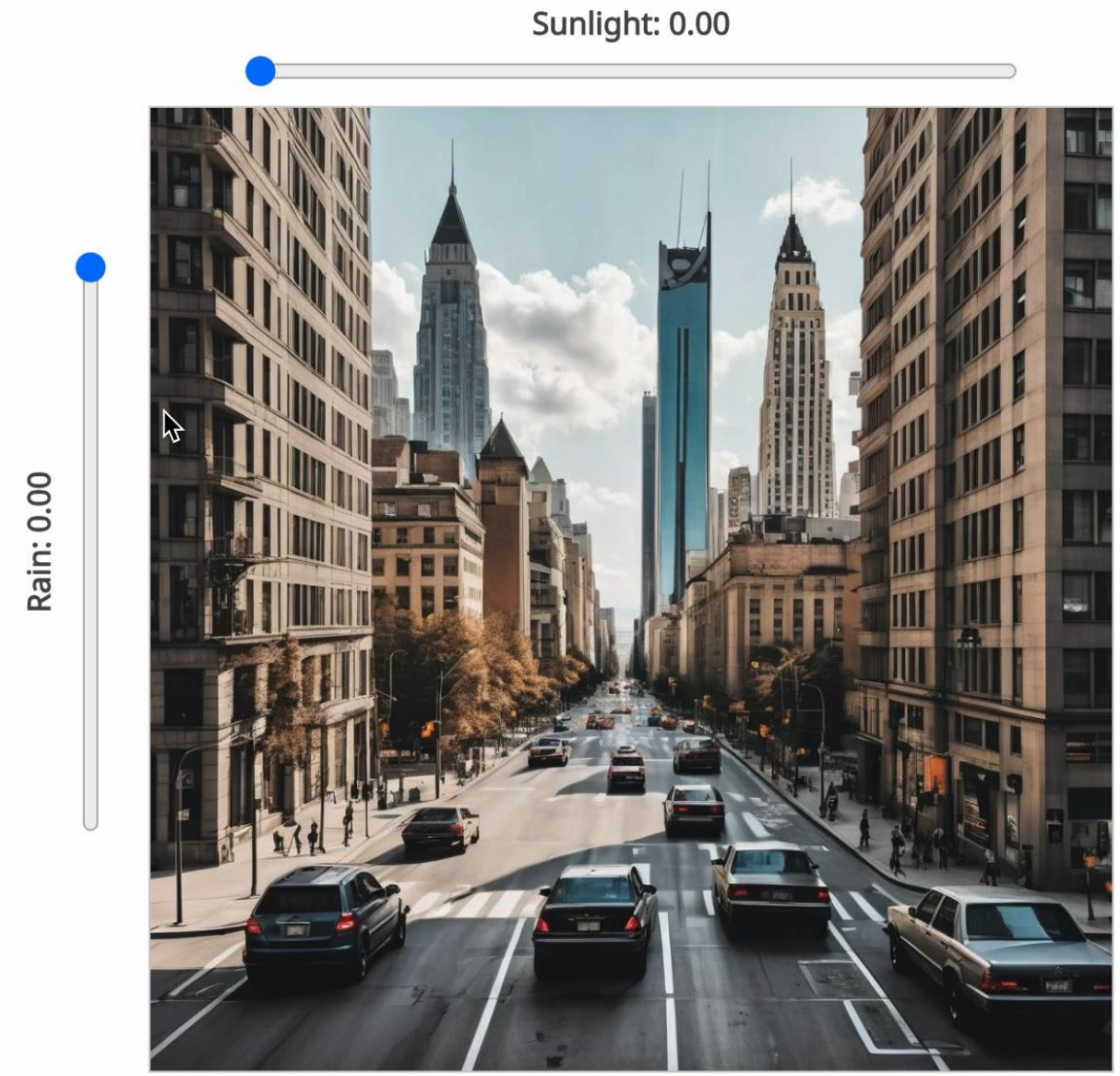
long exposure effects: 0.00

seaside

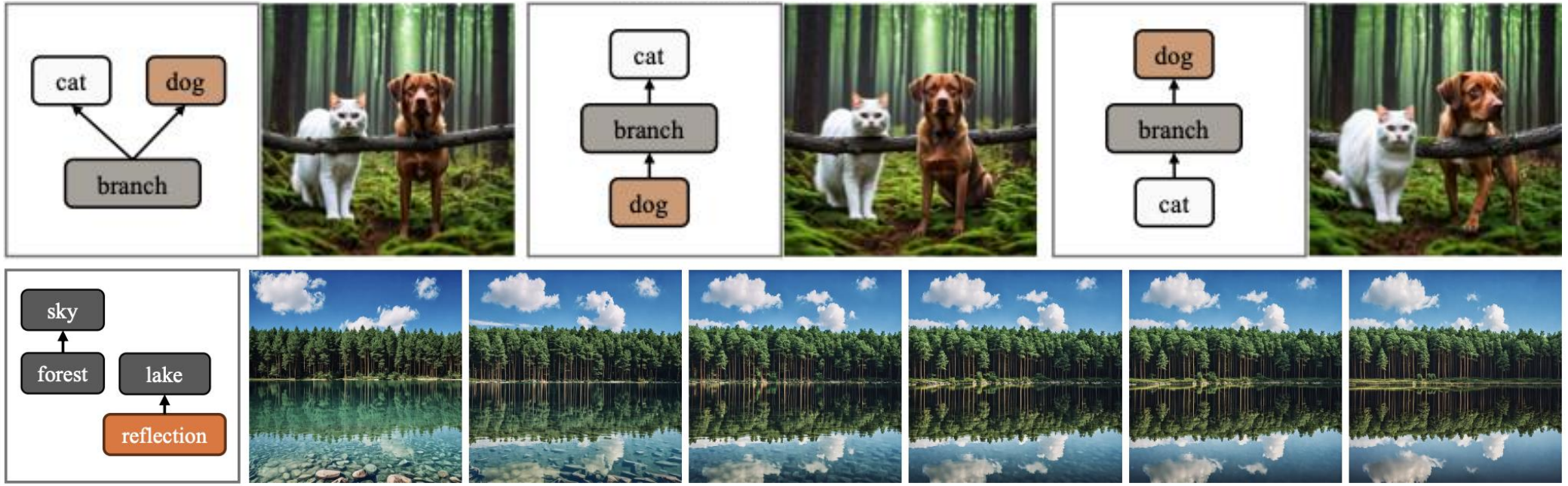
long exposure effect

Semantic opacity control





Q&A



LaRender provides training-free control over object occlusions and effect opacity in image generation.

Project page (code available)