

Accelerated Likelihood Maximization for Diffusion-based Versatile Content Generation

Anonymous ICLR 2026 submission

Overview: Various applications of our work

Image inpainting



Results with **unconditional diffusion model** trained on CelebA-HQ



Results with **Stable Diffusion XL**

Wide image generation



“Redwood forest, towering tranquility”



“Alpine village, snow-covered rooftops, nestled between majestic peaks—a picture-perfect scene of winter tranquility”

Long video generation

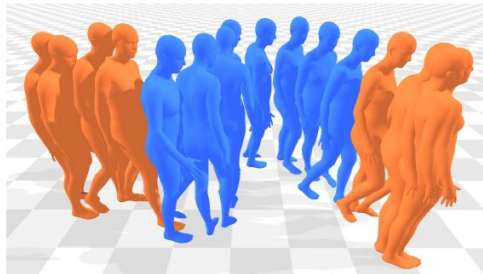


“Macro shot of bubbles rising in sparkling water, cinematic lighting, 8K.”



“Wide shot of waves shimmering under sunlight, high dynamic range, 4K clarity.”

Human motion inpainting



[Middle half] “the person walking back-and-forth in a zigzag.”

Introduction

- Limitations of the naive usage of diffusion models
 - Only generates fixed-size outputs
 - Limited generation from partially observed or pre-generated inputs
- Versatile content generation
 - Conditioning the generation process on partially observed inputs
 - Enables diffusion models to go beyond simple synthesis and instead infer missing or extended content
 - Key mechanism – inpainting & outpainting
- Still, developing a unified approach that generalizes across diverse content generation tasks remains an unsolved challenge.

Related Work

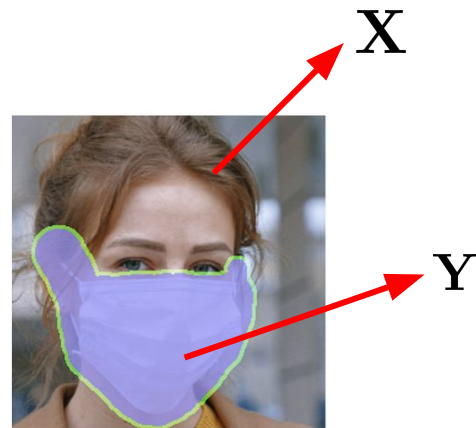
- Training-based approaches
 - Achieve superior task-specific performance
 - However, incur additional high computational training cost
 - Often struggle to generalize across domains
- Training-free approaches
 - Diffusion synchronization: a representative training-free strategy
 - *e.g.* SyncSDE (Lee et al., 2025): enforces consistency in observed regions while leaving unobserved regions unconstrained, assuming that realistic completions emerge from contextual alignment
 - However, often fails to produce plausible results (see the ‘w/o ALM’ column in the figure below)



Notation

- \mathbf{X} : observed content (unobserved parts have zero value)
- \mathbf{Y} : unobserved variable
- \mathbf{M} : binary mask indicating unknown regions as 1

- At diffusion timestep t ,
 - \mathbf{X}_t : noisy observed content
 - \mathbf{Y}_t : unobserved variable
 - $\mathbf{E}_t = \mathbf{X}_t + \mathbf{Y}_t \odot \mathbf{M}$



Preliminary: SyncSDE

- Conditional probability of unobserved variable given observed content:

$$p(\mathbf{X}_t \mid \mathbf{Y}_t, \mathbf{c}) := p(\mathbf{X}_t \mid \mathbf{Y}_t) \sim \mathcal{N}(\mathbf{Y}_t, w_1(1 - \alpha_t)(\mathbf{1} - \bar{\mathbf{M}})^{-1})$$

- Revised update equation of DDIM reverse process:

$$\mathbf{Y}_{t-1} = \sqrt{\alpha_{t-1}} \left(\frac{\mathbf{Y}_t - \sqrt{1 - \alpha_t} \epsilon_\theta(\mathbf{Y}_t, t, \mathbf{c})}{\sqrt{\alpha_t}} \right) + \sqrt{1 - \alpha_{t-1}} \epsilon_\theta(\mathbf{Y}_t, t, \mathbf{c}) + w_1(\mathbf{1} - \mathbf{M}) \odot (\mathbf{X}_t - \mathbf{Y}_t)$$

- Key limitations
 - Does not guarantee that the unobserved region will be harmonized with the observed content
 - Just assumes that synchronization will naturally produce a plausible outcome
 - Unobserved regions often contain inconsistent or unrealistic content
- Our solution: Accelerated Likelihood Maximization (ALM)

Proposed method: Accelerated Likelihood Maximization

- Optimization objective to minimize (with respect to $\Delta \mathbf{Y}_t^i$):

$$-\lambda_1 \log p(\mathbf{X}_t, \mathbf{M} \mid \mathbf{Y}_t^i + \mathbf{M} \odot \Delta \mathbf{Y}_t^i, \mathbf{c}) - \lambda_2 \log p(\mathbf{X}_t, \mathbf{M}, \mathbf{Y}_t^i + \mathbf{M} \odot \Delta \mathbf{Y}_t^i \mid \mathbf{c})$$

- Closed-form calculation

- Please check Section A of the Appendix for detailed derivation.

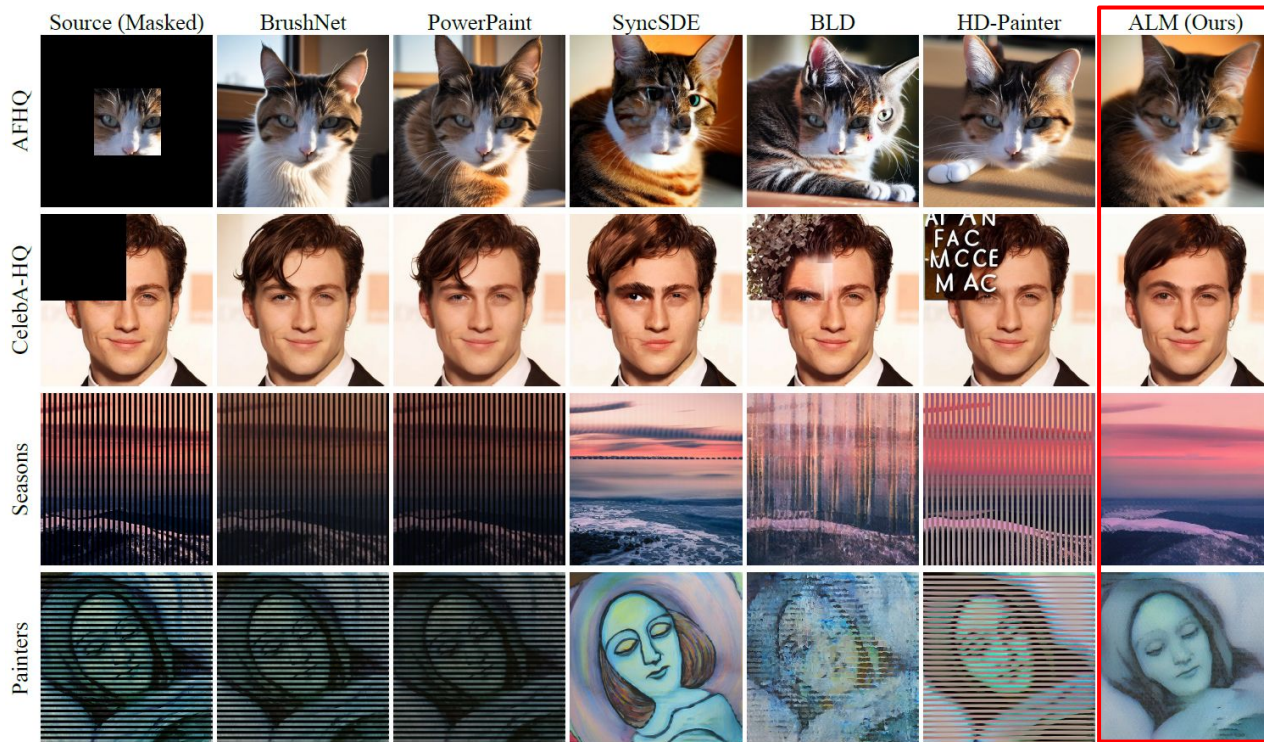
$$\begin{aligned} \Delta \mathbf{Y}_t^i &= \mathbf{M} \odot (\lambda_1 \nabla_{\mathbf{Y}_t^i} \log p(\mathbf{X}_t, \mathbf{M} \mid \mathbf{Y}_t^i, \mathbf{c}) + \lambda_2 \nabla_{\mathbf{Y}_t^i} \log p(\mathbf{X}_t, \mathbf{M}, \mathbf{Y}_t^i \mid \mathbf{c})) \\ &\longrightarrow \Delta \mathbf{Y}_t^i = \mathbf{M} \odot (\lambda_1 (\epsilon_\theta(\mathbf{Y}_t^i, t, \mathbf{c}) - \epsilon_\theta(\mathbf{E}_t^i, t, \mathbf{c})) - \lambda_2 \epsilon_\theta(\mathbf{E}_t^i, t, \mathbf{c})) \end{aligned}$$

- Revised DDIM reverse process:

$$\begin{aligned} \mathbf{Y}_{t-1} &= \sqrt{\alpha_{t-1}} \left(\frac{\mathbf{Y}_t - \sqrt{1 - \alpha_t} \epsilon_\theta(\mathbf{Y}_t, t, \mathbf{c})}{\sqrt{\alpha_t}} \right) + \sqrt{1 - \alpha_{t-1}} \epsilon_\theta(\mathbf{Y}, t, \mathbf{c}) \\ &\quad + w_1 (\mathbf{1} - \mathbf{M}) \odot (\mathbf{X}_t - \mathbf{Y}_t) + \mathbf{M} \odot (w_2 \epsilon_\theta(\mathbf{Y}_t, t, \mathbf{c}) - w_3 \epsilon_\theta(\mathbf{E}_t, t, \mathbf{c})) \end{aligned}$$

Results: Image inpainting

- Comparison with baselines



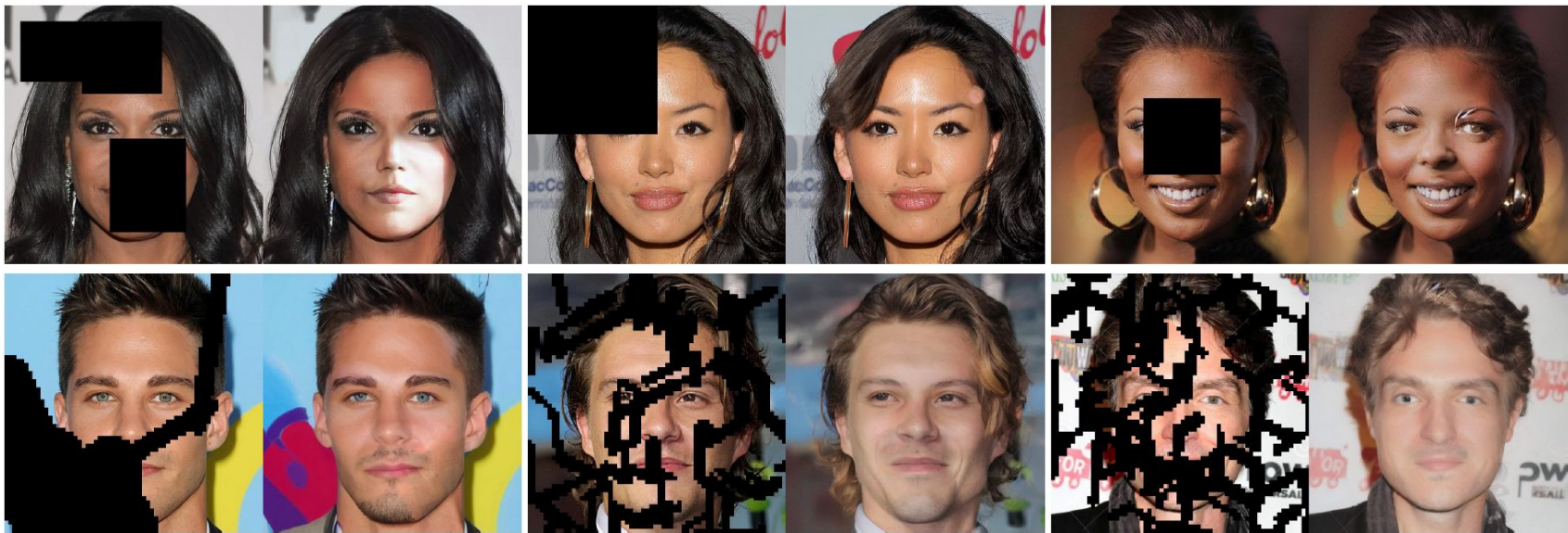
Results: Image inpainting

- Results on AFHQ dataset



Results: Image inpainting

- Results on CelebA-HQ dataset



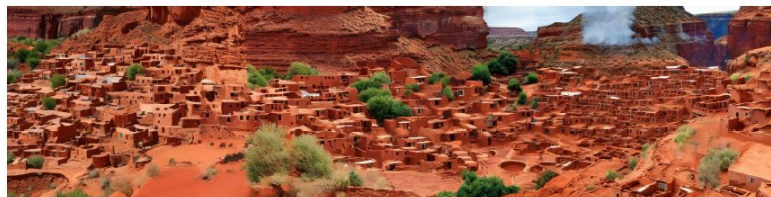
Results: Wide image generation

- Comparison with baselines

SyncTweedies



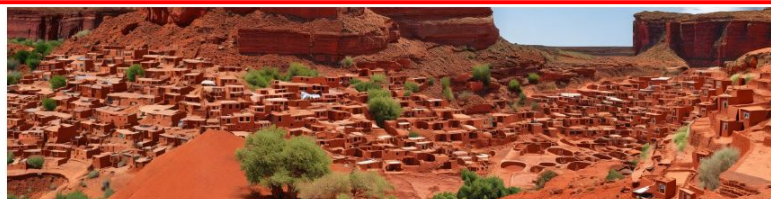
SyncSDE



ALM (Ours)



“Alpine village, snow-covered rooftops, nestled between majestic peaks—a picture-perfect scene of winter tranquility”



“Nestled in a canyon, a pueblo village stands against the red earth”

Results: Wide image generation

- Qualitative results using various prompts



“A photo of a forest with a misty fog”



“Alpine meadow, wildflowers swaying in a mountain breeze,
snow-capped peaks embracing a serene panorama
—a high-altitude sanctuary”



“Redwood forest, towering tranquility”



“Alpine village, snow-covered rooftops, nestled between majestic
peaks—a picture-perfect scene of winter tranquility”



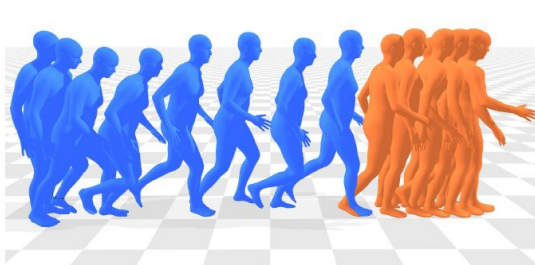
“Hidden waterfall, cascading down moss-covered
rocks in a tranquil glade”



“Inside a floating city above the clouds, suspended by levitating
platforms and connected by intricate sky bridges”

Results: Human motion completion

- First half prediction: **predict the initial part** of a sequence **given only the latter half**



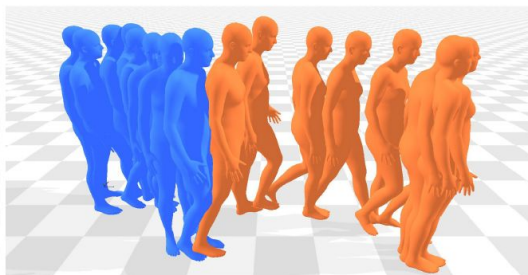
"he is running down then stopped and moved his right hand"



"person walks forward and stops"



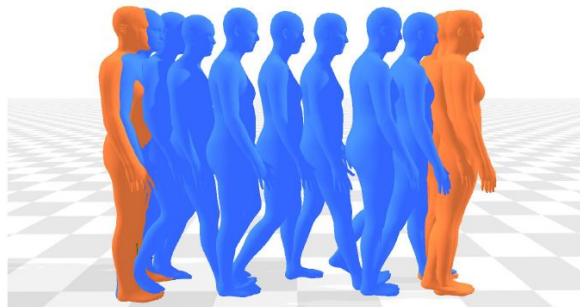
"a man takes several steps forward, jumps over an imaginary obstacle, lands on both feet, takes several more steps forward, turns around 180 degrees, takes several steps forward and jumps over same obstacle and returns to start."



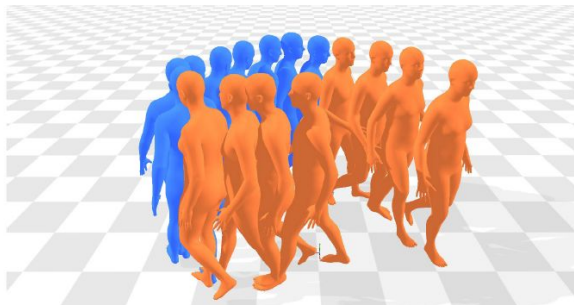
"the person is walking back-and-forth in a zigzag."

Results: Human motion completion

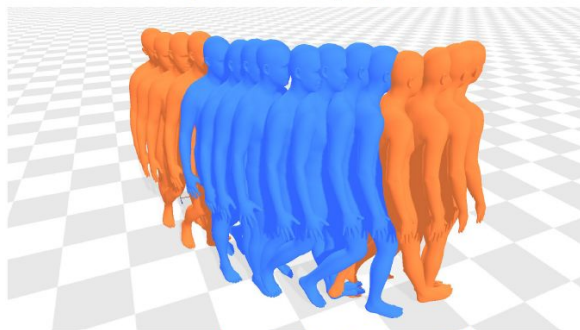
- Middle half completion: fill in the central portion given the first and last quarters



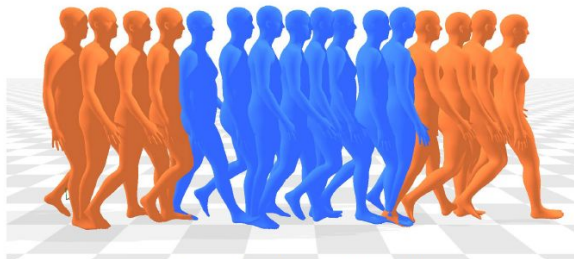
“the person is walking very slow”



“someone nervously pacing around in a circle”



“a man walks to the left side and stands.”



“forward walking and it ends”

Results: Long video generation

- Qualitative results: 16 \rightarrow 104-frame long video sequence



“Macro shot of smoke curling upward in the air, high contrast background, slow motion.”



“Macro shot of bubbles rising in sparkling water, cinematic lighting, 8K.”



“Wide shot of waves shimmering under sunlight, high dynamic range, 4K clarity.”

Results: Effects of ALM and its terms

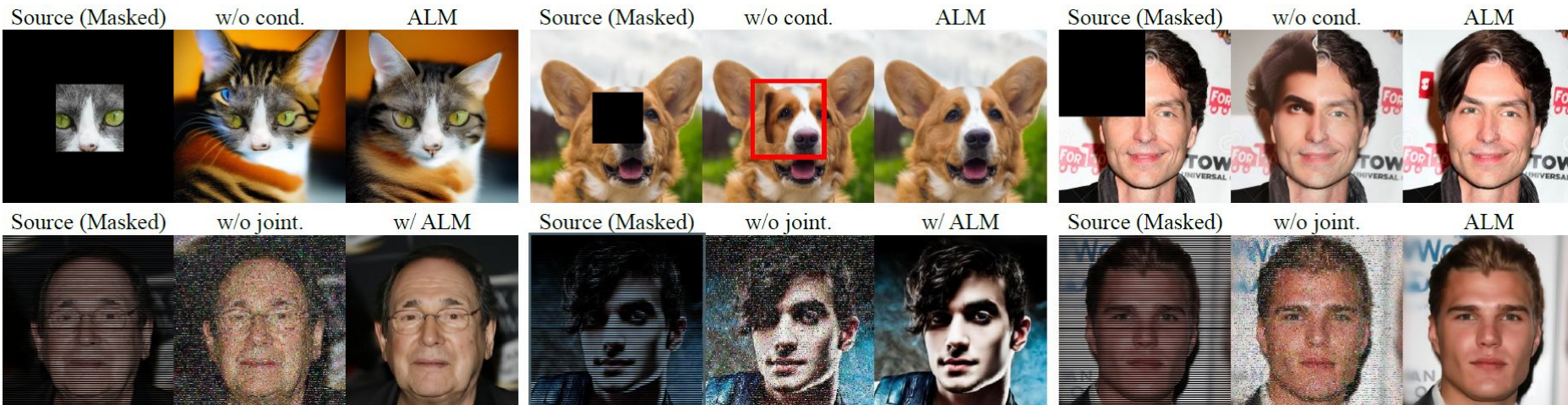
- Effects of ALM
 - ALM consistently improves unpromising baseline results and eliminates artifacts



Results: Effects of ALM and its terms

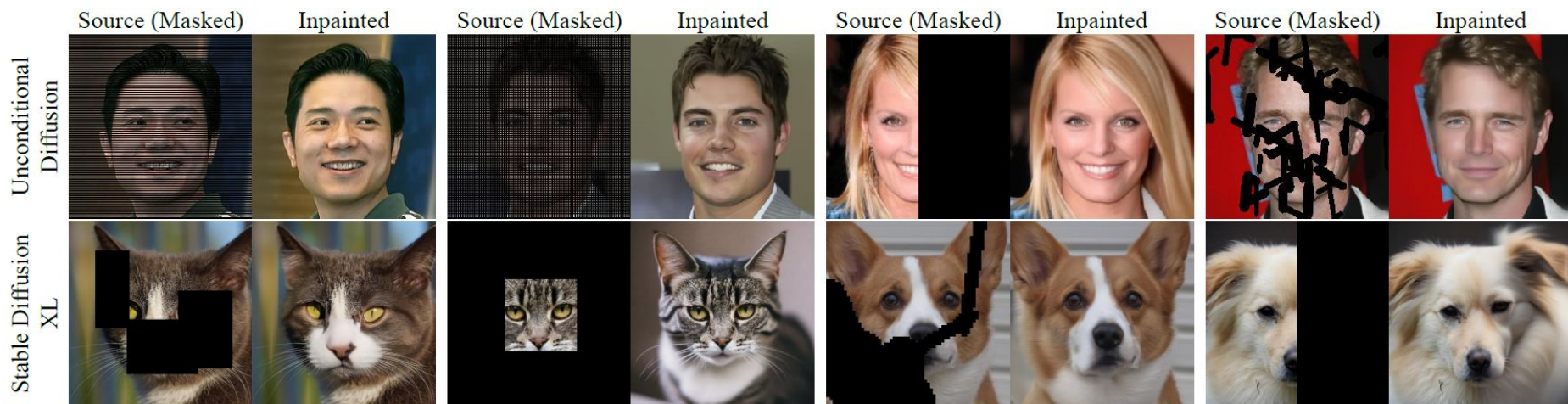
- Effects of conditional & joint likelihood terms in optimization objective
 - Both terms play a crucial role for high-quality output

$$\arg \min_{\Delta \mathbf{Y}_t} -\lambda_1 \log p(\mathbf{X}_t, \mathbf{M} \mid \mathbf{Y}_t^i + \mathbf{M} \odot \Delta \mathbf{Y}_t^i) - \lambda_2 \log p(\mathbf{X}_t, \mathbf{M}, \mathbf{Y}_t^i + \mathbf{M} \odot \Delta \mathbf{Y}_t^i)$$



Results: Robustness across diverse backbones

- Experiment using the unconditional diffusion model and SDXL
 - ALM shows superior performance on both backbones



Conclusion

- We introduce a novel sampling mechanism that explicitly models unobserved regions during diffusion sampling, overcoming the key limitation of prior synchronization methods.
- We establish a general training-free framework that extends across diverse modalities, including images, videos, and 3D human motion.
- We demonstrate the effectiveness of our approach through experiments on diverse tasks, achieving state-of-the-art performance compared to both training-based and training-free baselines.