VIPaint: Image Inpainting with Pre-Trained Diffusion Models via Variational Inference

Sakshi Agarwal, Gabriel Hope, Jimin Heo, Erik B. Sudderth

Accenture, Swarthmore College, Univ. California Irvine, Univ. California Irvine

Goal: Leverage diffusion models for realistic image inpainting.

Challenges for the state-of-the-art:

- **unstable / unreliable**
- **poor diversity**
- **large mask failures**

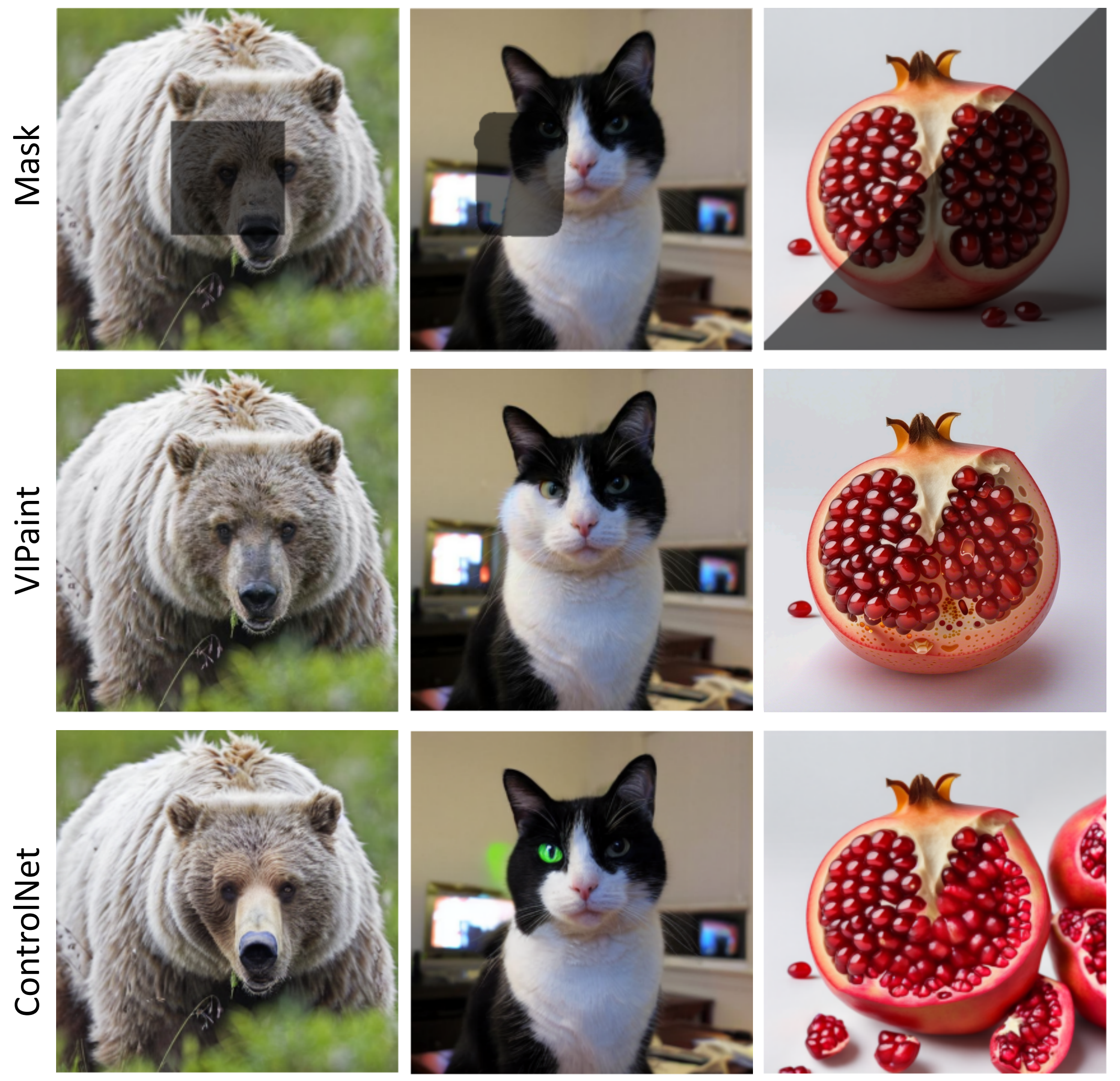
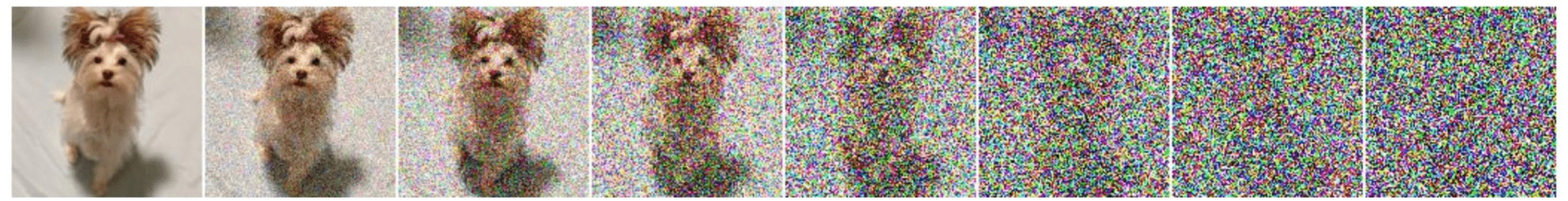

Mask
VIPaint
ControlNet

Image Inpainting by applying VIPaint to pre-trained Stable Diffusion 3.5, versus a baseline model retrained via ControlNet.

Background:



x Forward: Add noise Reverse: Denoise to reconstruct image z_T

Optimize VLB : $\mathcal{L}_{(0,T)}(z_0) = \frac{T}{2}\mathbb{E}_{\epsilon\sim\mathcal{N}(0,1),t\sim\mathcal{U}(1,T)}\left[||\epsilon - \hat{\epsilon}_\theta(z_t,t)||_2^2\right]$, $\hat{\epsilon}_\theta$ = denoiser

Inference : Given partial observation $y = x \odot m$, we want to recover true x

$$p_\theta(x \mid y) = \int_z p_\theta(z_T \mid y)p_\theta(x \mid z_0, y)\prod_{t=1}^{T} p_\theta(z_{t-1} \mid z_t, y)\, dz.$$

- Existing variational approach (**REDdiff** [Mardani ICLR23]) optimizes a *single* image (the mean of a low-variance Gaussian) that reconstructs y while simultaneously having high probability under the prior.
- **ControlNet** [Zhang ICCV2023] fine-tunes DMs with conditioning (e.g. masks/edges) to guide inpainting.

VIPaint:

Posterior : $q_\lambda(x, z \mid y) = p_\theta(x \mid z_{T_s})\left(\prod_{i=1}^{K-1} q_\lambda(z_{s(i)} \mid z_{s(i+1)})\right) q_\lambda(z_{T_e}),$

Optimize VLB : $L(\lambda) =$

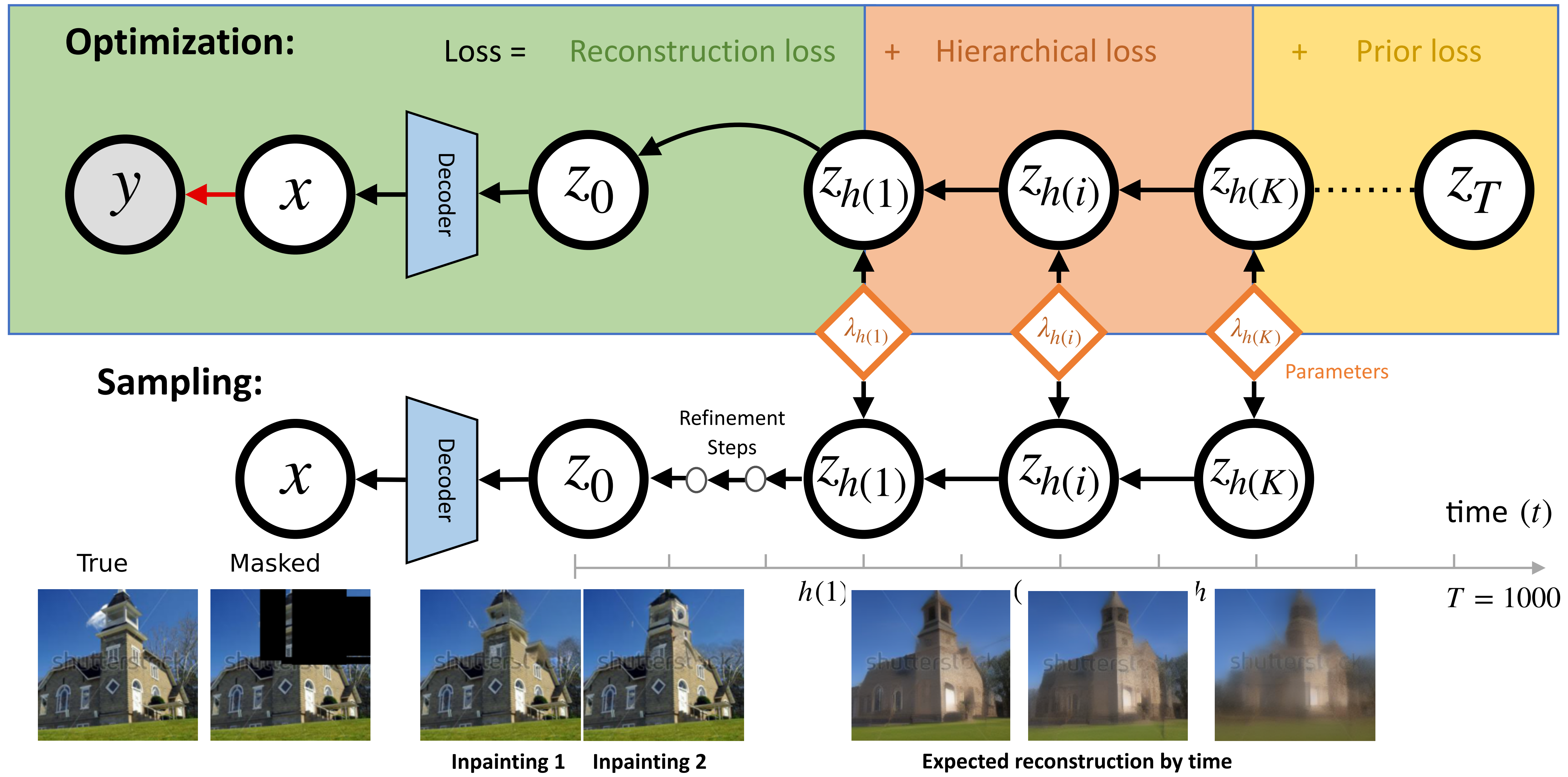
Reconstruction + Hierarchical + Prior

$$-\mathbb{E}_q[\log p_\theta(y \mid z_{T_s})] \quad \beta\sum_{i=1}^{K-1} D\left[q_\lambda(z_{s(i)} \mid z_{s(i+1)}) \,||\, p_\theta(z_{s(i)} \mid z_{s(i+1)}))\right] \quad \beta\mathcal{L}_{(T_e,T)}(z_{T_e})$$

→ matches observed pixels → stabilizes inference → aligns with diffusion model

Key takeaway: Hierarchical Latent Posterior

Instead of optimizing a single image or fine-tuning the prior diffusion model over a dataset, we infer a Markov posterior over latent variables for each observation.


Optimization:
Loss = Reconstruction loss + Hierarchical loss + Prior loss
y
x
Decoder
z_0
$z_{h(1)}$
$z_{h(i)}$
$z_{h(K)}$
z_T
$\lambda_{h(1)}$
$\lambda_{h(i)}$
$\lambda_{h(K)}$
Parameters
Sampling:
Refinement Steps
time (t)
$h(1)$
$T = 1000$
True
Masked

Inpainting 1 Inpainting 2

Expected reconstruction by time

Results : Pixel-based Diffusion Prior

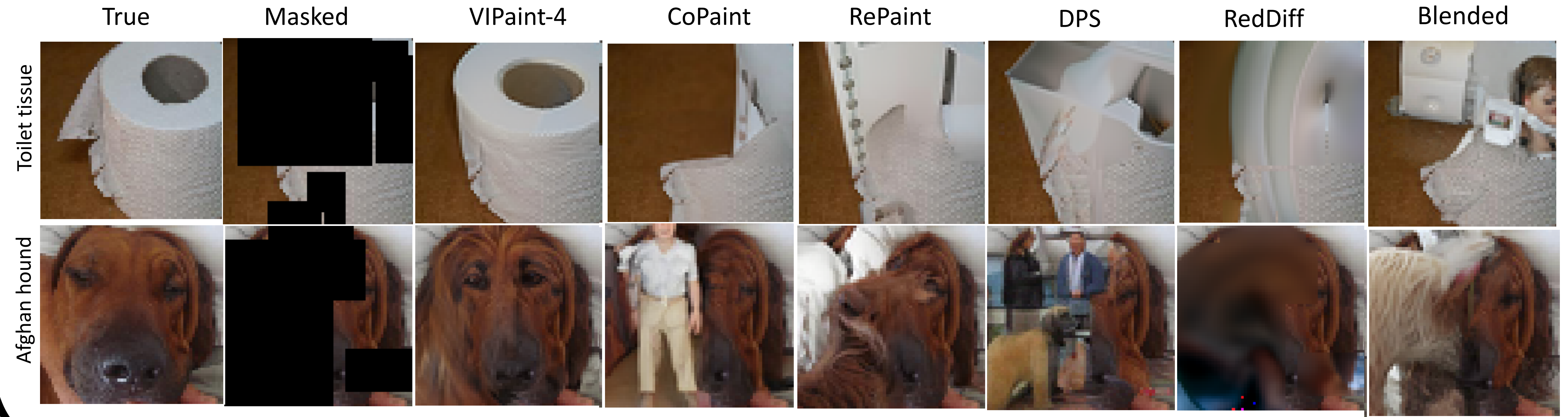
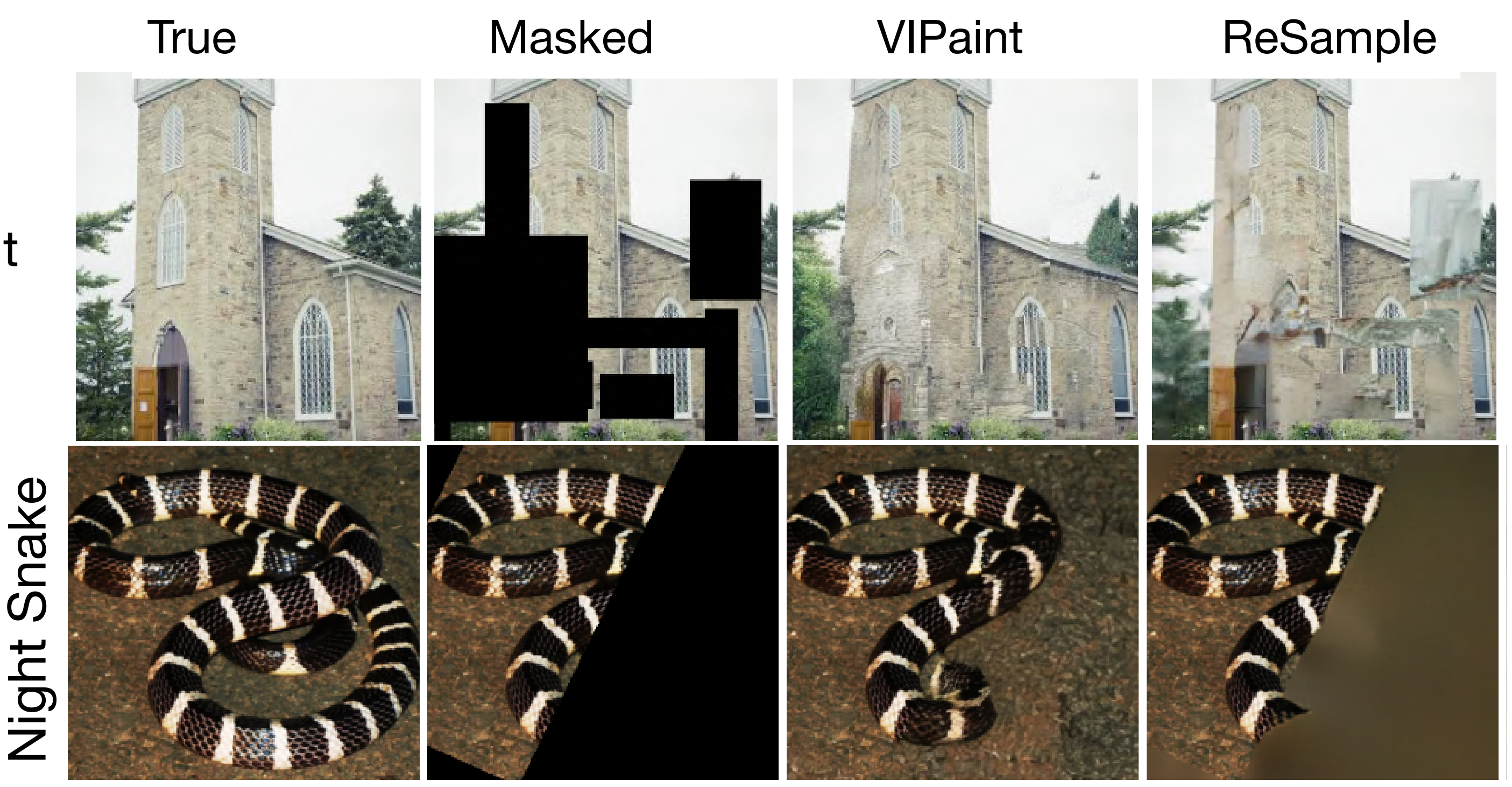

True
Masked
VIPaint-4
CoPaint
RePaint
DPS
RedDiff
Blended
Toilet tissue
Afghan hound

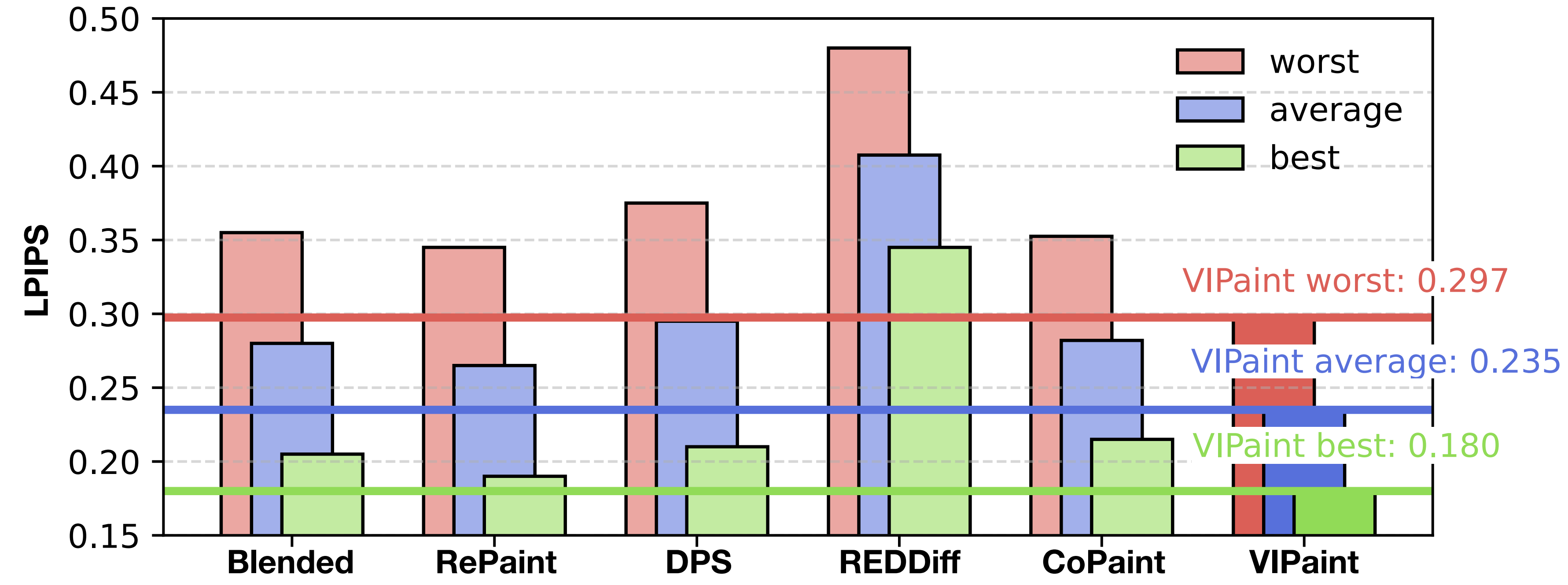
Image completion using pixel-based EDM model for ImageNet64 for large mask image inpainting. The best performing baseline, CoPaint, performs poorly.

Results : Latent Diffusion Prior

- ReSample gives blurry, inconsistent outputs
- VIPaint better preserves global structure, gives realistic images.

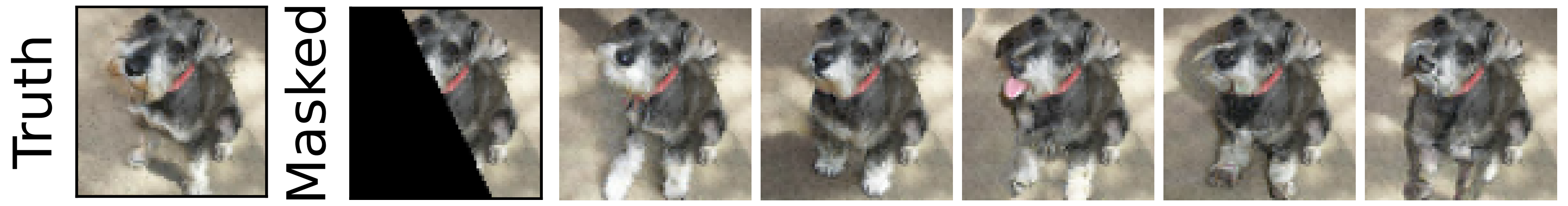
ImageNet,LSUN


True
Masked
VIPaint
ReSample
Night Snake

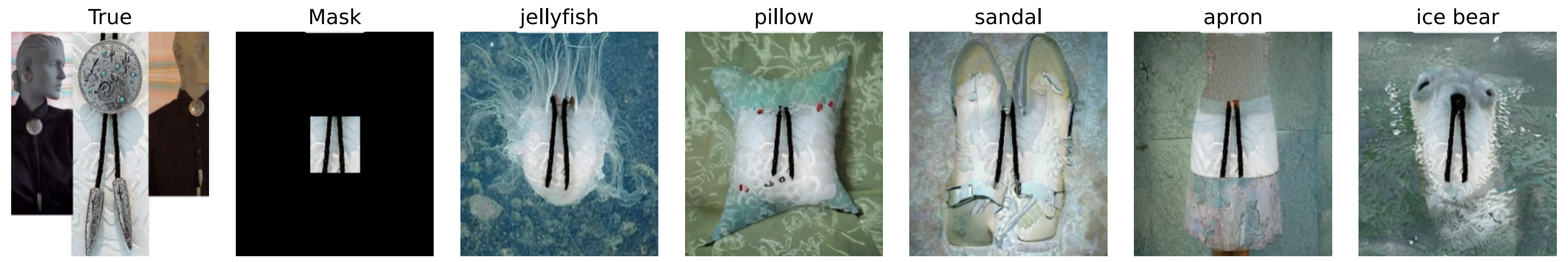


Here, LPIPs is estimated as the mean score of 1000 inpaintings with respect to the true image, averaged across the test set.

VIPaint captures multi-modal posteriors


Truth
Masked

5 Sample completions of VIPaint for a test images. We see that VIPaint yields coherent samples while capturing uncertainty in the missing pixels.


True
Mask
jellyfish
pillow
sandal
apron
ice bear

VIPaint diversity for ImageNet256 with LDM prior using different class conditioning. VIPaint follows the input label and ensures consistency.

Conclusion : **VIPaint provides stable and diverse inpaintings (and generalizes to other inverse problems, see paper) by modeling a hierarchical latent posterior.**

- Produces semantically consistent images
- Handles large mask scenarios
- Captures multi-modal posterior uncertainty

Acknowledgments: This research supported in part by NSF Robust Intelligence Award IIS-1816365, ONR Award N00014-23-1-2712, and by the HPI Research Center in Machine Learning and Data Science at UC Irvine.