A. ELBO derivation

We provide the full derivation of our model and losses from Equation (3). We start with our goal of finding model parameters θ that maximize the following probability for all videos and all t:

$$\begin{aligned} & p_\theta(\delta_t, x_{t-1}; x_T) \\ \propto & p_\theta(\delta_t | x_{t-1}; x_T) \\ = & \int_{z_t} p_\theta(\delta_t | z_t, x_{t-1}; x_T) p(z_t) dz_t. \end{aligned}$$

We use variational inference and introduce an approximate posterior distribution $q_\phi(z_t|\delta_t, x_{t-1}; x_T)$ [32, 63, 64].

$$\begin{aligned} & \int_{z_t} p_\theta(\delta_t | z_t, x_{t-1}; x_T) p(z_t) dz_t \\ = & \int_{z_t} p_\theta(\delta_t | z_t, x_{t-1}; x_T) p(z_t) \frac{q_\phi(z_t|\delta_t, x_{t-1}; x_T)}{q_\phi(z_t|\delta_t, x_{t-1}; x_T)} dz_t \\ \propto & \log \int_{z_t} p_\theta(\delta_t | z_t, x_{t-1}; x_T) p(z_t) \frac{q_\phi(z_t|\delta_t, x_{t-1}; x_T)}{q_\phi(z_t|\delta_t, x_{t-1}; x_T)} dz_t \\ = & \log \int_{z_t} \frac{p_\theta(\delta_t | z_t, x_{t-1}; x_T) p(z_t)}{q_\phi(z_t|\delta_t, x_{t-1}; x_T)} q_\phi(z_t|\delta_t, x_{t-1}; x_T) dz_t \\ = & \log \mathbb{E}_{z \sim q_\phi(z_t|\delta_t, x_{t-1}; x_T)} \left[\frac{p_\theta(\delta_t | z_t, x_{t-1}; x_T) p(z_t)}{q_\phi(z_t|\delta_t, x_{t-1}; x_T)} \right]. \end{aligned} \tag{8}$$

We use the shorthand $z_t \sim q_\phi$ for $z \sim q_\phi(z_t|\delta_t, x_{t-1}; x_T)$, and apply Jensen's inequality:

$$\begin{aligned} & \log \mathbb{E}_{z_t \sim q_\phi} \left[\frac{p_\theta(\delta_t | z_t, x_{t-1}; x_T) p(z_t)}{q_\phi(z_t|\delta_t, x_{t-1}; x_T)} \right] \\ \geq & \mathbb{E}_{z_t \sim q_\phi} \big[\log p_\theta(\delta_t | z_t, x_{t-1}; x_T) \big] \\ & + \mathbb{E}_{z \sim q_\phi} \big[\log \frac{p(z_t)}{q_\phi(z_t|\delta_t, x_{t-1}; x_T)} \big] \\ \geq & \mathbb{E}_{z_t \sim q_\phi} \big[\log p_\theta(\delta_t | z_t, x_{t-1}; x_T) \big] \\ & - KL[q_\phi(z_t|\delta_t, x_{t-1}; x_T) || p(z_t)], \end{aligned} \tag{9}$$

where $KL[\cdot||\cdot]$ is the Kullback-Liebler divergence, arriving at the ELBO presented in Equation (5) in the paper.

Combining the first term in Equation (5) with our image likelihood defined in Equation (1):

$$\begin{aligned} & \mathbb{E}_{z_t \sim q_\phi} \log p_\theta(\delta_t | z_t, x_{t-1}; x_T) \\ \propto & \mathbb{E}_{z_t \sim q_\phi} \big[\log e^{-\frac{1}{\sigma_1^2} |\delta_t - \hat{\delta}_t|} \\ & + \log \mathcal{N}\big(V(x_{t-1} + \delta_t); V(x_{t-1} + \hat{\delta}_t), \sigma_2^2 \mathbb{I}\big) \big] \\ = & \mathbb{E}_{z_t \sim q_\phi} \left[-\frac{1}{\sigma_1^2} |\delta_t - \hat{\delta}_t| \right. \\ & \left. + \log \frac{1}{\sqrt{2\pi\sigma_2^2}} \exp\big(- \frac{(V(x_{t-1} + \delta_t) - V(x_{t-1} + \hat{\delta}_t))^2}{2\sigma_2^2} \big) \right] \\ \propto & \mathbb{E}_{z_t \sim q_\phi} \left[-\frac{1}{\sigma_1^2} |\delta_t - \hat{\delta}_t| \right. \\ & \left. - \frac{1}{2\sigma_2^2} (V(x_{t-1} + \delta_t) - V(x_{t-1} + \hat{\delta}_t))^2 \right], \end{aligned} \tag{10}$$

giving us the image similarity losses in Equation (6). We derive $\mathcal{L}_{KL}$ in Equation (6) by similarly taking the logarithm of the normal distributions defined in Equations (2) and (4).

B. Network architecture

We provide details about the architecture of our recurrent model and our critic model in Figure 11.

C. Human study

We surveyed 150 human participants. Each participant took a survey containing a training section followed by 14 questions.

Calibration: We first trained the participants by showing them several examples of real digital and watercolor painting time lapses.

Evaluation: We then showed each participant 14 pairs of time lapse videos, comprised of a mix of watercolor and digital paintings selected randomly from the test sets. Although each participant only saw a subset of the test paintings, every test painting was included in the surveys. Each pair contained videos of the same center-cropped painting. The videos were randomly chosen from all pairwise comparisons between real, *vdp*, and *ours*, with the ordering within each pair randomized as well. Samples from *vdp* and *ours* were generated randomly.

Validation: Within the survey, we also showed two repeated questions comparing a real video with a linearly interpolated video (which we described as *interp* in Table 2 in the paper) to validate that users understood the task. We did not use results from users who chose incorrect answers for one or both validation questions.

D. Additional results

We include additional qualitative results in Figures 12 and 13. We encourage the reader to view the supplementary video, which illustrates many of the discussed effects.

We examine failure cases from the proposed method in Figure 14, such as making many fine or disjoint changes in a single time step and creating an unrealistic effect. We show the effect of increasing the number of samples k in Figure 15.

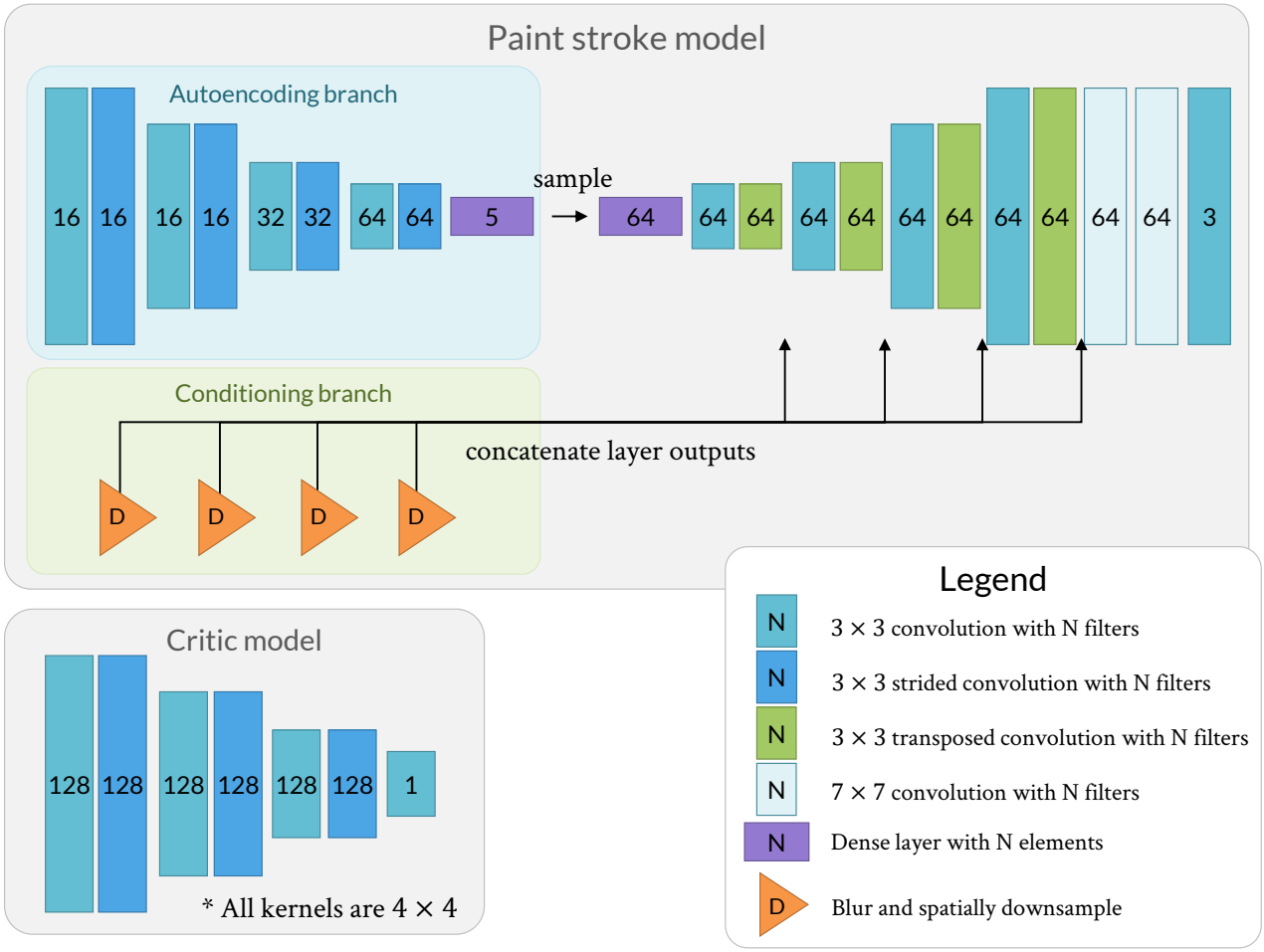
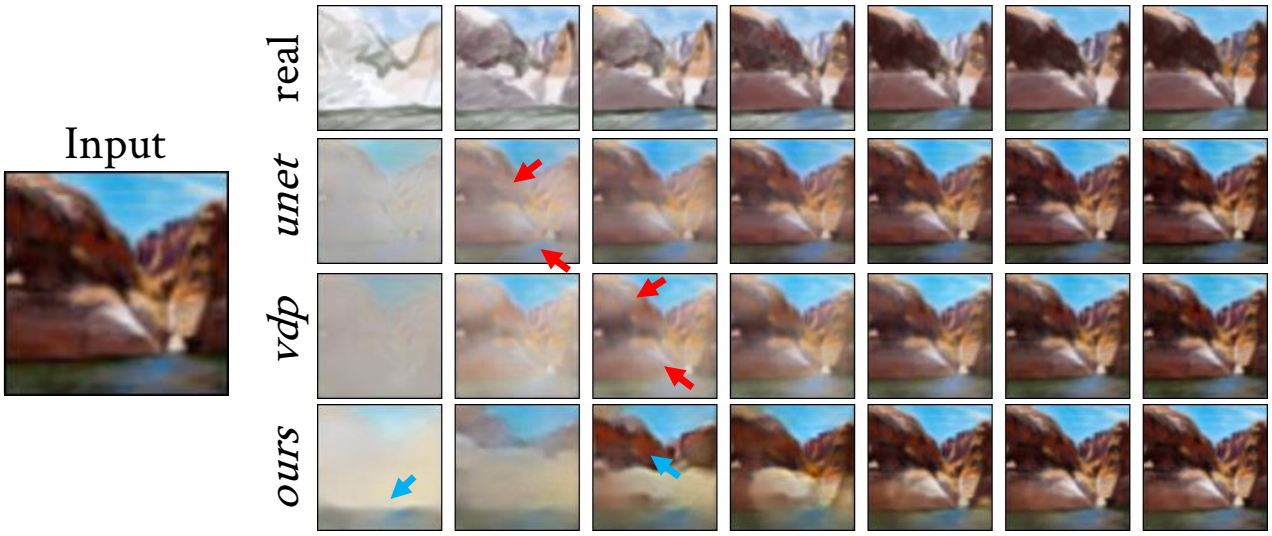
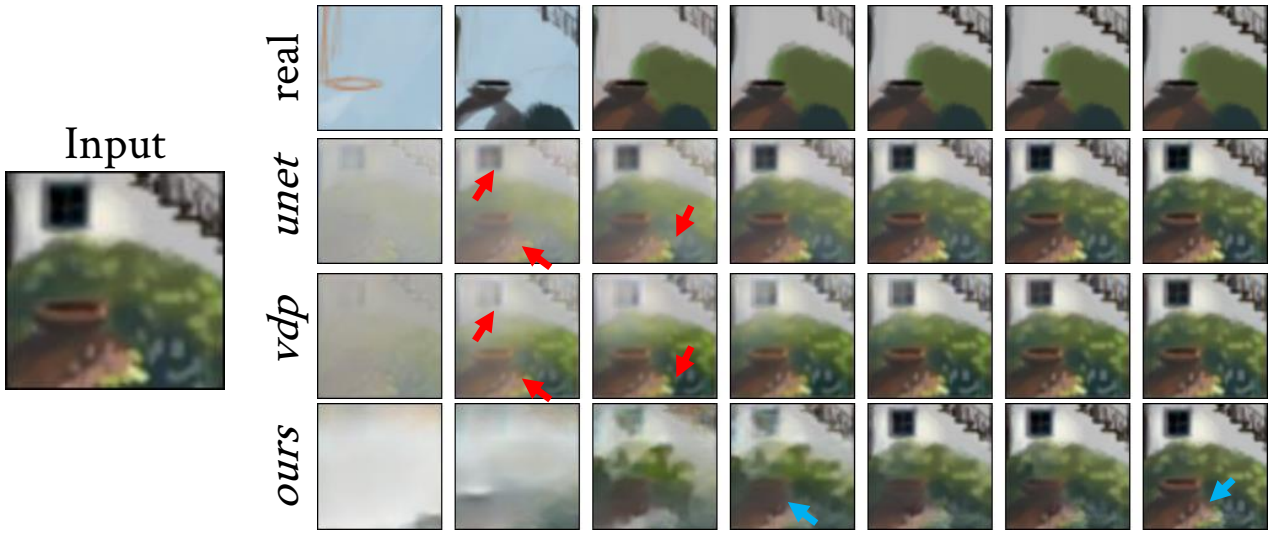


Figure 11: **Neural network architecture details**. We use an encoder-decoder style architecture for our model. For our critic, we use a similar architecture to StarGAN [10], and optimize the critic using WGAN-GP [19] with a gradient penalty weight of 10 and 5 critic training iterations for each iteration of our model. All strided convolutions and downsampling layers reduce the size of the input volume by a factor of 2.

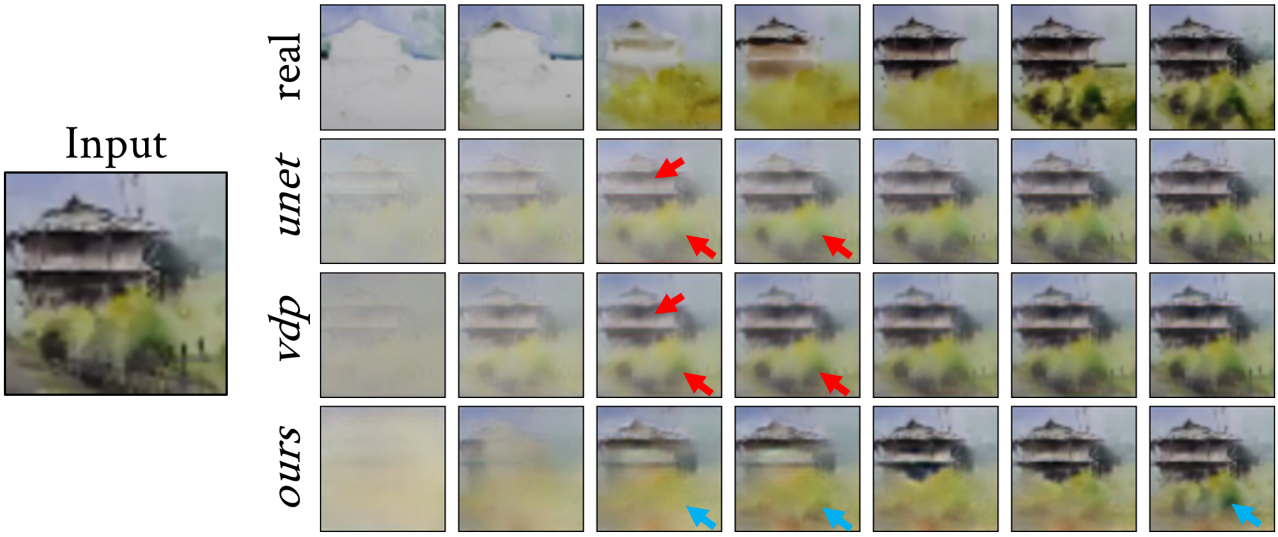


(a) **The proposed method paints similar regions to the artist**. Red arrows in the second row show where *unet* adds fine details everywhere in the scene, ignoring the semantic boundary between the rock and the water, and contributing to an unrealistic fading effect. The video synthesized by *vdp* produces more coarse changes early on, but introduces an unrealistic-looking blurring and fading effect on the rock (red arrows in the third row). Blue arrows highlight that our method makes similar *painting changes* to the artist, filling in the base color of the water, then the base colors of the rock, and then fine details throughout the painting.

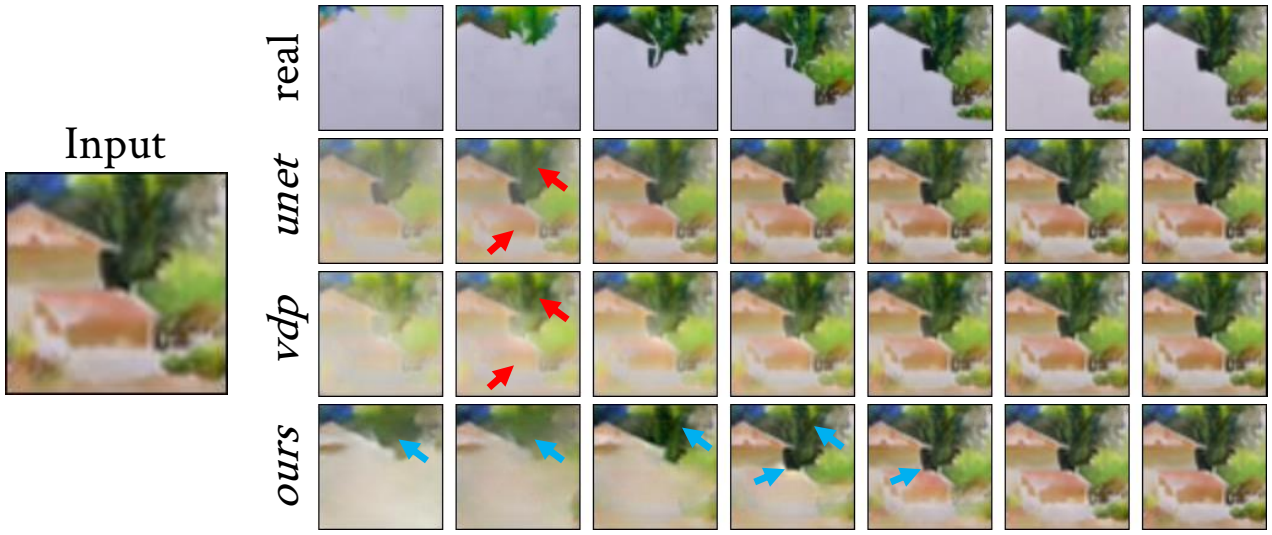


(b) **The proposed method identifies appropriate colors and shape for each layer of paint**. Red arrows indicate where the baselines fill in details that the artist does not complete until much later in the sequence (not shown in the real sequence, but visible in the input image). Blue arrows show where our method adds a base layer for the vase with a reasonable color and shape, and then adds fine details to it later.

Figure 12: **Videos synthesized from the watercolor paintings test set**. For the stochastic methods *vdp* and *ours*, we examine the nearest sample to the real video out of 2000 samples. We discuss the variability among samples from our method in Section 5, and in the supplementary video.

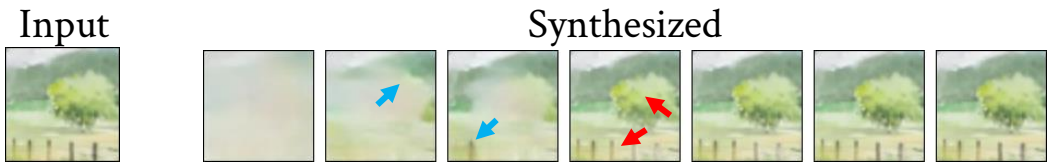


(a) **The proposed method paints using coarse-to-fine layers of different colors, similarly to the real artist**. Red arrows indicate where the baseline methods fill in details of the house and bush at the same time, adding fine-grained details even early in the painting. Blue arrows highlight where our method makes similar *painting changes* to the artist, adding a flat base color for the bush first before filling in details, and using layers of different colors.



(b) **The proposed method synthesizes watercolor-like effects such as paint fading as it dries**. Red arrows indicate where the baselines fill in the house and the background at the same time. Blue arrows in the first two video frames of the last row show that our method uses coarse changes early on. Blue arrows in frames 3-5 show where our method simulates paint drying effects (with the intensity of the color fading over time), which are common in real watercolor videos.

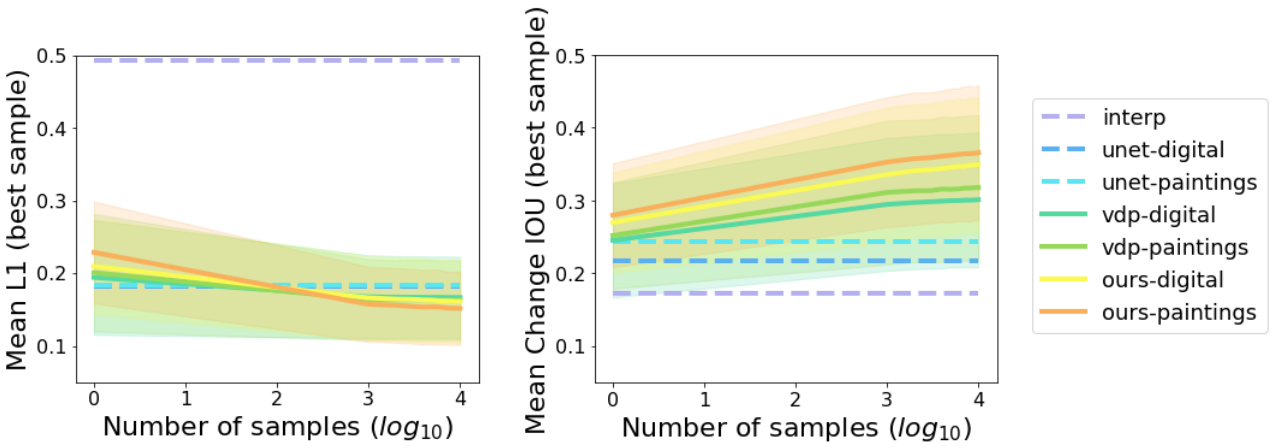
Figure 13: **Videos synthesized from the watercolor paintings test set**. For the stochastic methods *vdp* and *ours*, we show the nearest sample to the real video out of 2000 samples.



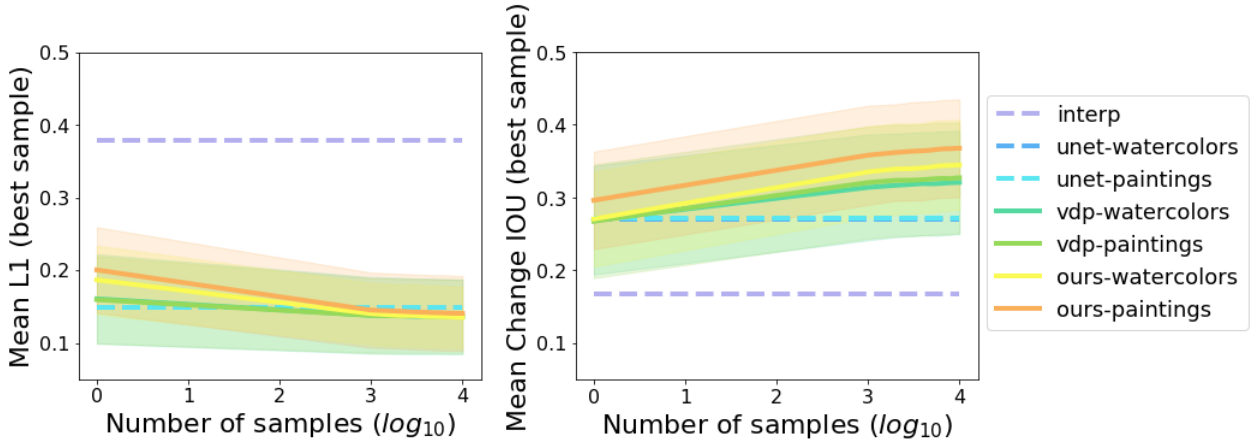
(a) **The proposed method does not always synthesize realistic changes for fine details**. Blue arrows highlight frames where the method makes realistic painting changes, working in one or two semantic regions at a time. Red arrows show examples where our method sometimes fills in many details in the frame at once.

(b) **The proposed method sometimes synthesizes changes in disjoint regions**. Red arrows indicate where the method produces painting changes that fill in small patches that correspond to disparate semantic regions, leaving unrealistic blank gaps throughout the frame. This example also fills in much of the frame in one time step, although most of the filled areas in the second frame are coarse.

Figure 14: **Failure cases**. We show unrealistic effects that are sometimes synthesized by our method, for a watercolor painting (top) and a digital painting (bottom).



(a) Digital paintings test set.



(b) Watercolor paintings test set.

Figure 15: **Quantitative measures**. As we draw more samples from each stochastic method (solid lines), the best video similarity to the real video improves. This indicates that some samples are close to the artist's specific painting choices. We use L1 distance as the metric on the left (lower is better), and change IOU on the right (higher is better). Shaded regions show standard deviations of the stochastic methods. We highlight several insights from these plots. (1) Both our method and *vdp* produce samples that are comparably similar to the real video by L1 distance (left). However, our method synthesizes painting changes that are more similar in shape to those used by artists (right). (2) At low numbers of samples, the deterministic *unet* method is closer (by L1 distance) to the real video than samples from *vdp* or *ours*, since L1 favors blurry frames that average many possibilities. (3) Our method shows more improvement in L1 distance and painting change IOU than *vdp* as we draw more samples, indicating that our method captures a more varied distribution of videos.