We thank the reviewers for their insightful comments. We are encouraged that reviewers found our paper clear, very well written, well motivated (**R1,R3,R4**). In particular, **R2** thinks the idea of explicitly modeling missingness is important. **R3** finds the missingness latent variable and dynamic appearance to be novel. **R4** recognizes the effectiveness of our method in reconstructing the missing information in the input video as well as predicting for future frames. Below we address specific concerns, and we will incorporate all feedback in the final version.

R1, R3 *... exact same model architecture *without* the missingness* We consider DDPAE to be the closest architecture *without* the missingness variable. We also test the exact same model architecture with and without missingness variable for Scenario 2 of Moving MNIST, as shown in the table:

Mov. MNIST Scenario 2	BCE ↓		MSE ↓		PSNR ↑		SSIM ↑	
Model (trained 300 epochs)	Rec	Pred	Rec	Pred	Rec	Pred	Rec	Pred
DIVE w/o missingness	236.35	356.82	49.07	76.52	19.40	17.66	0.86	0.82
DIVE w missingness	**165.42**	**321.29**	**27.03**	**64.17**	**22.15**	**18.56**	**0.93**	**0.83**

R2 *Line 153 says that the model is trained maximizing a variational lower bound...* Our model strictly follows the VAE framework (see caption of Fig. 1) by maximizing the ELBO objective as below:

$$\begin{aligned}\log p_\theta(\mathbf{y}^{1:K}, \mathbf{x}^{K+1:T}) \quad &\geq \quad \mathbb{E}_q\left[\log p_\theta\left(\mathbf{y}^{1:K}|\mathbf{z}_{1:N}^{1:K}\right) - KL\left(q_\phi\left(\mathbf{z}_{1:N}^{1:K}\right)||p\left(\mathbf{z}_{1:N}^{1:K}\right)\right)\right] \\ &+ \quad \mathbb{E}_q\left[\log p_\theta\left(\mathbf{x}^{K+1:T}|\mathbf{z}_{1:N}^{K+1:T}\right) - KL\left(q_\phi\left(\mathbf{z}_{1:N}^{K+1:T}\right)||p\left(\mathbf{z}_{1:N}^{K+1:T}\right)\right)\right]\end{aligned}$$

We will include these details in the main paper. Each component of $\mathbf{z}_i^t = [\mathbf{z}_{i,a}^t, \mathbf{z}_{i,p}^t, \mathbf{z}_{i,m}^t]$ is modeled separately.

R2 *...details on how the model is trained.* Due to space limits, we deferred the training and datasets details to the supplemental material (supp.). We also included source codes to reproduce our results.

R2, R3 *... the Heavyside function is not differentiable.* The Heavyside function $H(x)$ is **not** a variable node in our computational graph, hence we are NOT differentiating through it. We use *torch.where()* function in Pytorch to implement this condition operation. As shown in Fig. 1 top and Eq. (10), $\mathbf{z}_{i,m}$ is a masking indicator. In practice, we multiply each decoded object by the **logit** before the Heavyside function instead of the binary label (see Supp. L406 and L407). Hence, $\mathbf{z}_{i,m}$ gets its gradients from the decoder. We will clarify this in the updated version.

R1, R2 *...knowing the number of objects .* The number of objects N is specified a priori, but only as an upper bound. For the MOTS dataset, we set $N = 3$ but the actual number of objects is often lower. Similar to DDPAE, our model can learn to set the redundant components to be empty.

R2 *What kind of distribution is the one in Equation (5)...Is there a prior distribution?* As stated in L121, Eq. (5) is the variational distribution. We use Gaussian as prior distributions, parametrized by mean and variance (L125, L138). The specific values for these parameters are provided in the code.

R3 *there is no ablation study ... measure the impact of dynamic appearance vs static.* We **do** have ablation study in the Supp. due to space constraints. We **did** compare the impact of dynamic appearance modeling vs static in Supp., exactly as suggested by the reviewer.

R2, R3 *How is the mixture p set in practice (eq (2))?...* For Eq. (2) we set $p = 0.25$ and for Eq. (8), $p = 0.85$ (see supp. L402-L404). We experimented with different values of p (± 0.1) and did not find any significant difference.

R2 *talks about covariate shift but this is never elaborated on ...* The covariate shift comes from the distributional difference between training and testing data. Without the mixing strategy in Eq. (8), the model would overfit to the dynamic component of the appearance during training. Our random mixing regularizes the model to learn a better representation for static appearance.

R3 *... In fig 4 and 5 ...results reported in the DDPAE's paper, images are much worse, Why?* Because our tasks are significantly more **difficult** than the experiments of DDPAE's paper. Fig. 4 and 5 show objects out-of-scene and dynamic appearance. DDPAE assumes that there is no missing object and that the appearance of each object remains static. Missing objects hinder the tracking and appearance modeling for DDPAE. If an object's appearance varies over time, DDPAE will learn an average appearance, leading to blurry reconstructions and errors in tracking.

R1 *DRNet is not a stochastic video prediction method.* This is a typo. DRNet is a strong unsupervised video prediction baseline as it also learns disentangled representations from video.

R4 *to be seen how well it can perform in more complex datasets* We appreciate your recognition of our contributions! We will consider more complex datasets and scenarios as future work, as also suggested by **R1**.