We thank the reviewers for their constructive comments on our paper. We address the major questions in the following.

R1: The ability to handle large temporal inconsistency. Our approach can handle large temporal inconsistency such as multimodal inconsistency with the IRT strategy and outperform baselines, as shown in Table 2, Table 3, Fig. 5, and the supplement. In the experiments for most evaluated tasks, we do not observe the extreme cases mentioned by R1. As for the extreme case, we believe no existing blind temporal consistency approach can perform well. In the colorization example, note that red color is an inconsistent artifact and is thus learned as a minor case (see Fig. 6). We do not observe obvious blurry artifact in the mentioned CycleGAN case and users prefer our results compared with the original video.

R1: The performance degradation problem. Avoiding performance degradation is critical to blind temporal consistency, and perceptual similarity is used to analyze performance degradation. Table 2, Table 3, Fig. 5, and the supplement show quantitative and qualitative comparisons. For dehazing, the comparison with original videos is provided at $0:13$ in the supplement and our performance is not degraded perceptually. The user study takes both perceptual preservation and temporal consistency into account, and the preference rate of our method is much higher than that of baselines.

R1: Just an extension of DIP? Our DVP is not a direct extension of DIP, and there are substantiate differences in several aspects (1) Implementation. DIP reconstructs the image from noise while DVP tries to learn the mapping from input frames to processed frames. A simple extension of DIP on video should be: reconstructing the video from noise by a 3D-CNN. (2) Assumption. DIP assumes image prior exists in CNN architectures; DVP assumes video consistency enforced by correspondences can be learned from the internal similarity of frames. (3) Application. DVP can enable numerous image processing methods applicable to videos while maintaining temporal consistency. (4) IRT. The proposed IRT solves the multimodal inconsistency problem well, which is ignored by prior work. We treat flickering artifacts of unimodal inconsistency as noises in the temporal domain, which is an important observation.

R1, R2, R4: Train a specific network for each video and running time. Training on a test video takes about 2 seconds per frame, which is not real-time. Moreover, the comparison of 1000 times is not reasonable since test environment is different. Compared with direct inference, extra 24 / 49 epochs are required for training. Besides, we can try to speed up the model by using a lighter model. Also, our approach has advantages: no need for training on a large dataset, which may take hours or even days; domain gap between a training set and a test set does not exist.

R1, R2, R3, R4: Carefully tuning? Relationship with the length of videos. How to select the epochs? Is it still "Blind"? (1) One basic observation is that reconstructing the flickering artifacts takes **much more** time compared with common video contents. For example, in Fig. 9, E_{warp} is only increased by 0.002 from 25-th to 80-th epoch. Hence, we do not need to tune the epochs carefully. (2) Since the network learns a temporal consistent image mapping, the training time is decided by the iterations. For the same kind of video, a video with 200 frames requires fewer epochs than a video with 50 frames. It is an interesting idea to further study the relationship between video length and the number of epochs needed. (3) In the experiments, since the temporal consistency is great in many epochs, we simply select the same epoch (25 or 50 epochs) for all the videos (30 to 200 frames) in a task based on a small validation set up to 5 videos. (4) Our approach is blind because we always treat every image operator as a black box.

R1, R3, R4: Other metrics, e.g., LPIPS, VGG loss or original metrics. We use PSNR because PSNR is a common metric for data fidelity in many tasks we evaluate. We also have a user study to evaluate human perceptual preference. We agree that LPIPS and VGG loss are good metrics for perceptual similarity, and we will add VGG loss and LPIPS in the final version. Fig. 9 should be the curve of perceptual similarity and temporal inconsistency mentioned by R4.

R2, R4: 3D CNN, distillation and optical flow. Great advice on more comparisons. (1) Reconstructing a video from noises by a 3D CNN can be a simple extension of DIP. Hence, we believe such a 3D CNN model is memory hungry, and we will analyze it. Also, the multimodal problem can be a challenge in this simple method. (2) Our method works for both learning and non-learning based operators. If the image operator is a CNN, it is similar to train a student network from the teacher network on a single video. We believe we can adopt a lighter CNN architecture to speed up. (3) Using optical flow is often useful for short-term temporal consistency. However, optical flow is usually not accurate enough for long-term consistency (Line 34-35, caption in Fig. 5), and more comparison is described at Line 72-87.

R3, R4: Clarification for IRT in Sec. 3.2. With IRT, we increase the number of channels in the network output (e.g., six channels for two RGB images). Then, in each iteration, Eq. 5 is used to divide pixels into two clusters (main and minor modes). This is similar to cluster assignment in K-Means when K=2 and the pixels in minor mode are similar to the outliers in IRLS. We use $L1$ distance as spatial distance d, as mentioned in Line 146-147. Then pixels in two clusters are used for updating two modes in each iteration. The two modes are generally different in different iterations since they are obtained from different pixels. At last, notations will be revised, thank you.

R4: Discussion and exposition. We will add discussion and revise Fig. 2, the paragraph from Line 272, and Sec. 3.2.

R1, R2, R4: Additional related work. R1: Yes, these methods use the metric and we will include them. R2: We will analyze these works and add them. R4: Yes, this work also uses some type of video prior and we will discuss it.