

A ADDITIONAL ABLATIONS

A.1 IP ADAPTER VS CAT

A popular method for image-conditioning in image diffusion models is IP Adapter [Ye et al. \(2023\)](#). A CNN feature extractor takes the conditional input views and extracts features that will then be added to the intermediate features in the diffusion model forward pass. Here we compare it to using the conditioning method from CAT3D that directly uses conditional inputs as frames to the diffusion model without noising. We generally find that they are similar but IP adapter can exhibit more abrupt transitions between the input condition and neighbouring regions in the generated panorama. We show a few examples in [Fig. A1](#).

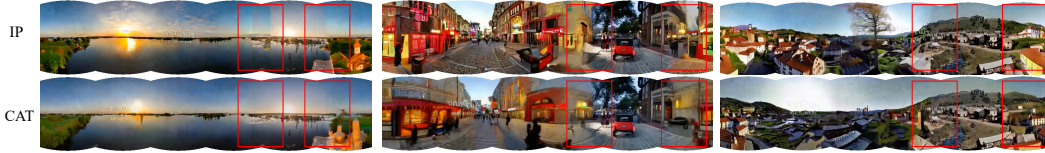


Figure A1: Qualitative figure comparing IP vs CAT type architecture for input conditioning. When using IP adapter, the consistency between input conditioning views and neighbouring views (highlighted in red box) is worse compare to CAT.

A.2 ABLATING THE EFFECTS OF SHIFTING THE NOISE SCHEDULE

During inference we use up to $8 \times 16 = 128$ frames which is much larger than the 16 frames used by the base video model. As mentioned in [§ 3.1](#) the increased data dimensionality also requires a corresponding increase in terminal noise to minimize the terminal step gap with the noise prior. In particular we interpolate between the standard noise schedule and a noise schedule that has been shifted by 10. We compare these qualitatively in [Fig. A2](#). Note that without changing the noise schedule, the model is largely incapable of generating plain regions such as clear sky or white snow fields and instead fills in the frame with visual clutter.

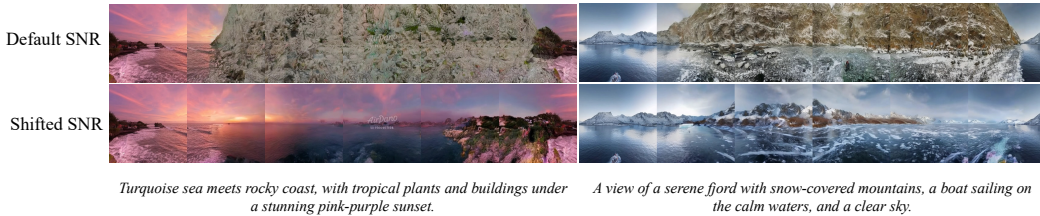


Figure A2: Qualitative comparison of shifting the noise schedule in the video-conditioned setting. Each of the six horizontal views is visualized independently before stitching into a panorama. Without shifting toward higher noise levels, the model struggles to generate clear skies or water, introducing objects that disrupt scene cohesion (e.g., sudden mountains and rocks).

A.3 ABLATING THE EFFECT OF NOISE AUGMENTATION FOR AUTOREGRESSIVE GENERATION

In this section, we qualitatively analyze the impact of noise augmentation during training on the model’s autoregressive generation performance. To demonstrate this, we compare two models: one trained with noise augmentation and the other without. To maximize the effect of error accumulation, we use both models to 6 frames at a time, for a total of 10 iterations to get a video consisting of $10 \times 5 + 1 = 51$ frames. [Figure A3](#) shows a side-by-side comparison of the different scenarios.

As autoregressive iterations increase, the model without noise-augmentation produces increasingly saturated frames. While the model trained with noise augmentation also shows some degradation, it maintains significantly better output quality over time, demonstrating its usefulness in reducing the severity of error accumulation.

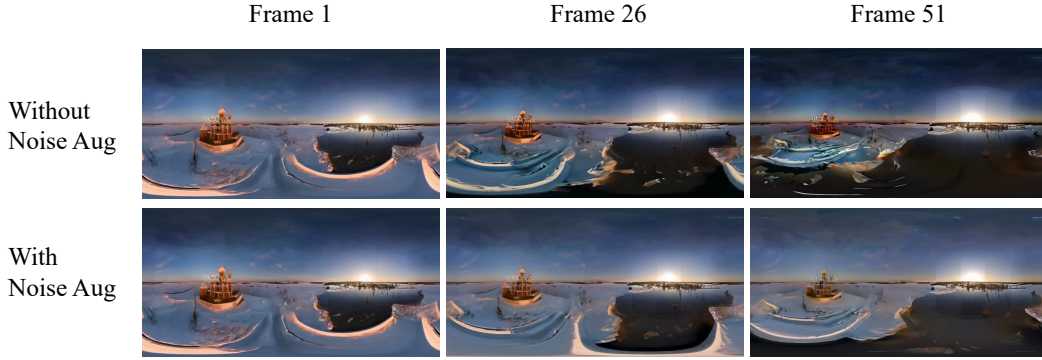


Figure A3: Qualitative comparison of autoregressive generation with and without noise augmentation. Both models exhibit a decline in output quality over time, but the model trained without noise augmentation shows a more rapid and severe degradation, with frames becoming increasingly saturated. In contrast, the model with noise augmentation deteriorates more gradually.

A.4 ABLATING THE EFFECTS OF FREEZING BASE MODEL LAYERS



Figure A4: Qualitative figure comparing text conditional panorama video generation using base model freezing vs no freezing. Freezing model weights is better able to retain some of the prior knowledge on out of distribution prompts.

When finetuning our model for multi-view generation we choose to freeze the base model layers. We ablate this choice qualitatively here on the text conditional panorama video generation task. We evaluate out of distribution prompts that make the overfitting behaviour very obvious when not freezing any base layers as can be seen in Fig. A4.

B ADDITIONAL VIDEO CONDITIONAL RESULTS

We show more video conditional generation results in Fig. B1 where we also apply autoregressive generation to extend the video length.

C ADDITIONAL TRAINING DETAILS

During the first stage of training we adapt the base video model towards the shifted and interpolated noise schedule as well as the v-prediction parameterization. This stage is trained for 10,000 iterations on the original dataset and a batch size of 128. Following that we insert the multi-view attention layers and train our model using the multiview video data. The batch size for this phase is 32 and we train these models for 15,000 iterations. Both stages use a constant learning rate of 0.0001. Most of our experiments are conducted on 32 A100 GPUs (or lower using gradient accumulation).

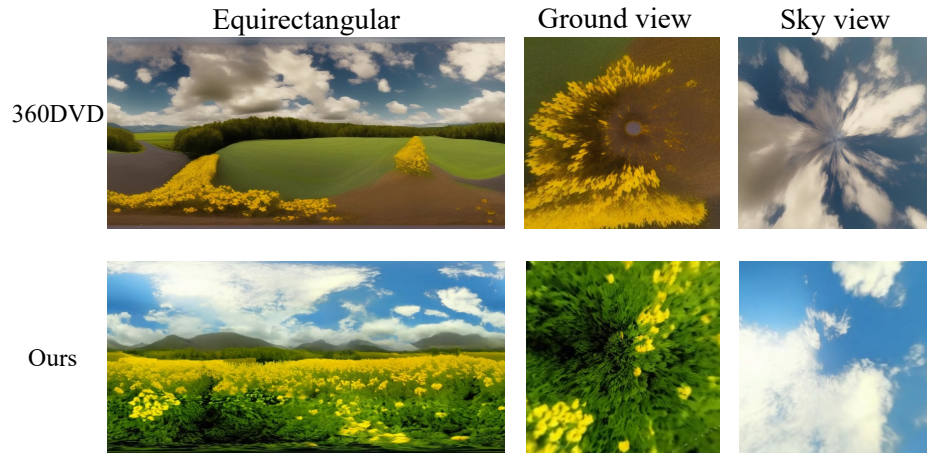


Figure A5: Qualitative figure comparing text conditional video generation, 360DVD VS ours and highlighting the distortion in 360DVD near the poles. Note that both generations were first transformed to the same equirectangular format before consistent sky and ground views were extracted. 360DVD struggles in these views as the distortion is highest here and deviates the most from perspective view images whereas we natively generate perspective views.



Figure B1: More results of video conditional autoregressive generation on out of distribution videos.