Sound2Sight: Generating Visual Dynamics from Sound and Context

Moitreya Chatterjee*[1] Anoop Cherian[2]

[1] University of Illinois at Urbana-Champaign, Urbana IL 61801, USA
[2] Mitsubishi Electric Research Laboratories, Cambridge MA 02139, USA
metro.smiles@gmail.com cherian@merl.com

1 Introduction

In the supplementary materials, we include the following:

1. Additional details about the three datasets and the associated tasks in Sections 2.1, 2.2, 2.3, and 2.4 respectively.
2. The architecture of *Sound2Sight* and details of the training procedure.
3. Standard Deviation measures of the model performance.
4. Auxiliary evaluation of the quality of the generated videos.
5. Performances from ablative studies of our model, the effect of the choice of hyper-parameters, per-sample comparisons with competing methods, and a study of the effectiveness of teacher forcing training strategy. Also included are plots showcasing the diversity of our model. Further, the alignment of the generated frames against the input audio is quantitatively evaluated.
6. Qualitative experimental results vis-á-vis competitive baselines and figures illustrating the diversity of the samples generated by our model.
7. Failure cases of our model.
8. **Video samples are also available in the supplementary zip file** and contain prototypical samples from the three different datasets and results produced by our scheme. We also provide qualitative video generation comparisons to competing methods, while showcasing our model's capability for diverse generation. We also include a video file showcasing the synchronization of the generated frames against the input audio. There are **eight** video files in the zip file: they are:
   - Sample_Moving_MNIST.avi: A sample clip from the M3SO dataset.
   - Sample_AudioSet_Drums.mp4: A sample clip from the AudioSet Drums dataset.
   - Sample_Painting.mp4: A sample clip from the Youtube Painting dataset.
   - M3SO_NB_5_15_Sample_Results.mp4: Sample generations of our method, vis-á-vis competing methods on the M3SO-NB dataset. This file also features diverse generations by our method.
   - M3SO_30_30_Sample_Results.mp4: Sample generations of our method, vis-á-vis competing methods on the M3SO dataset. This file also features diverse generations by our method.

* Work done as an intern at MERL.

- Drums_15_15_Sample_Results.mp4: Sample generations of our method, vis-á-vis competing methods on the Audioset-Drums dataset. This file also features diverse generations by our method.
- Painting_15_15_Sample_Results.mp4: Sample generations of our method, vis-á-vis competing methods on the Youtube Painting dataset. This file also features diverse generations by our method.
- Synchronized_Video_Final.mp4: This file reflects the synchronization of the generated video with the input audio for samples from the M3SO dataset. Besides visualizing the frames of the generated video, we also present the evolution of SSIM scores of the generated frames with time.

Furthermore, we include three folders containing raw output samples of our method for the M3SO, Audioset-Drums, and Youtube Painting datasets respectively. Accordingly, the folders are named: (i) Sound2Sight_Generated_M3SO_Clips; (ii) Sound2Sight_Generated_Drums_Clips; (iii) Sound2Sight_Generated_Painting_Clips. VLC player is the **recommended** player for the videos. **Also, note that all the videos, except the raw output videos in the 3 folders mentioned above, have audio as well. So kindly unmute your speakers.**

2 Datasets and Tasks

As described in the main paper, we present results on three Audio-Visual datasets for our video generation task, namely (i) the Multimodal Moving MNIST with (and without) a Surprise Obstacle (M3SO), and its variant without the obstacle (M3SO-NB), (ii) the Audioset-Drums [3], and (iii) YouTube-Painting. Here, we provide more details of these tasks. **Please see the supplementary video samples for a better understanding of the task**. Figure 2 shows samples from all three datasets and Table 1 presents the statistics of these datasets and the training/val/test splits.

2.1 Multimodal Moving MNIST with a Surprise Obstacle (M3SO)

This is a novel synthetic dataset, where a randomly chosen MNIST digit [6] moves in a box (of size 48×48), along rectilinear trajectories, in the seen part of the sequence (30 frames for example). While doing so, the digit occasionally bounces against the walls of the box which alters its trajectory of motion randomly. The digit then traverses this path of motion linearly till the next collision happens. Additionally, we associate audio with each digit (a unique tone for the digit), and the amplitude of this tone is inversely proportional to the distance of the moving digit from the lower-left corner of the box. When the digit bounces off the box edges, a different tone is emanated (i.e. the frequency of the audio changes). The audio then reverts to the original tone of the digit as it continues its motion. This process sustains as we transition from the "seen" frames to the ensuing "unseen" part.

At a pre-selected frame in the "unseen" part (the 42^{nd} frame, i.e. the 12^{th} unseen frame) , a square obstacle is introduced in the box at a randomly chosen

spatial location. When the digit bounces off this obstacle yet another different but unique tone is played (i.e. the frequency changes again). Post the collision, the digit continues its linear motion in a random direction, with its accompanying tone switching back to the original tone of the digit.

Given this setting, our task is to use sound to generate the unseen frames which involves capturing the dynamics of the digit, and also placing the obstacle accurately both in time and space. For our experiments, we train all algorithms with 30 seen frames and the task is to predict the next 30 for training. Additionally, in order to evaluate the generalizability of the model into the distant future, we ask all methods to predict 60 unseen frames at test time. The obstacle is introduced in the 42^{nd} frame, i.e. the 12^{th} unseen frame. Figure 1 represents this setup visually. Figure 2 shows some sample frames from a clip in this dataset.

2.2 M3SO-NB

We also conduct experiments on the proposed Multimodal Moving MNIST dataset without the surprise obstacle component (M3SO-NB). In this setting, we train with 5 seen frames and predict the next 15 frames. However at test time all competing models predict 25 frames.

2.3 Audioset-Drums

This dataset was constructed by collecting videos from the *Drums* class of the AudioSet dataset [3]. This particular class of videos is unique in the sense that most of the videos in this class have correlated visual and auditory information. We selected those videos from this class which clearly had some body-parts (mostly hands and head) of the drummer visible - playing his (or her) drum kit in an indoor environment. We removed videos that had animations, and those clips in which the sound source (i.e., the drum kit) was not clearly visible. All video clips were resized to a frame resolution of 64×64 at 30fps and the audio was sampled at 44kHz. Figure 2 shows some sample frames from this dataset. For this dataset, we train all competing techniques with 15 seen frames and predict the next 15. At test time however, we predict further into the future by predicting the next 30 frames after the 15 seen frames.

2.4 YouTube-Painting

Apart from the constrained or simplified audio-visual prediction context, as captured in the previous two datasets, we decided to introduce a more challenging dataset, that is still constrained in its context, however is diverse and loose in its spatio-temporal dynamics. After looking at a tradeoff between various possibilities on a large collection of Youtube videos, including diversity in camera view angles, types of actor motions, kinds of visual context, and a clear and distinctive audio cue, we decided to use videos containing an actor painting some art. We realized that there exists a good collection of such videos on Youtube, which

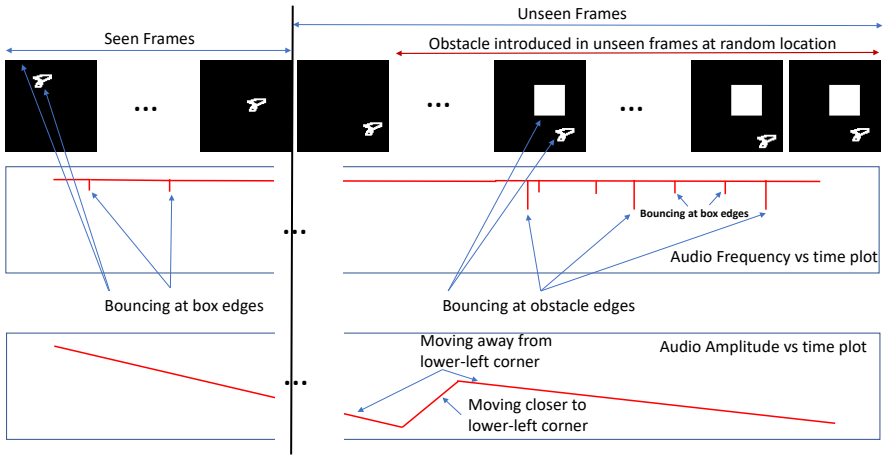


Fig. 1: An illustration of our proposed "Multimodal Moving MNIST with a Surprise Obstacle" (M3SO) dataset. The top row shows a few salient frames in a sample video. The two rows below it show the frequency (middle row) and amplitude (bottom row) of the audio signal plotted against time. When the digit bounces on the edges, it has a momemtary tone change, seen on the left part of the middle row, and when the digit bounces against the block, there is a different tone played mometarily, as is shown by the first spike on the right side of the middle row. The last row shows the change in amplitude of the tone as the digit moves closer/farther away from the bottom left corner of the canvas.

can provide a good training set for our scheme. While, this dataset may not be characteristic of large motions, it contains videos taken from multiple viewing angles, periodic motions (while painting), diversity in the visual context (such as different mixes of colors, paint brushes, etc.), different painters, and different orientations and types of painting canvas, brush, etc at the same time containing clear sound of the painter's brush touching the canvas. With this idea, we present our Youtube-Paining dataset.

The YouTube-Painting dataset was constructed by collecting videos from YouTube where a painter (or a part of his/her body) is seen painting an acrylic painting, inside a room [1]. We used "Taylor ASMR" as the search query to crawl these videos. Besides the visuals of the painter painting, there is accompanying audio emanating as a result of the painter's brush movements upon the canvas. The videos do not feature too much camera motion and thus the change in audio frequency is a cue for the distance of the brush from the camera. Akin to the AudioSet-Drums dataset, the chosen videos in this dataset also do not contain animations, shot changes, etc. Further, videos for which the sound source, i.e., the brush, was not clearly seen, were dropped from the dataset. Video clips in this dataset have frames of size 64×64 at 30fps, while the audio is sampled at

Table 1: A summary of the statistics of the different datasets.

Datasets	# Train	# Test	# Val	# Frames/video
M3SO	8000	1000	1000	100
AudioSet-Data	6000	1000	1000	90
YouTube-Painting	4800	500	500	90

44kHz. Figure 2 shows sample frames from this dataset. Here too, we train our models with 15 seen frames and predict the next 15. At test time however, we predict further into the future by predicting the next 30 frames after having seen the first 15 frames.

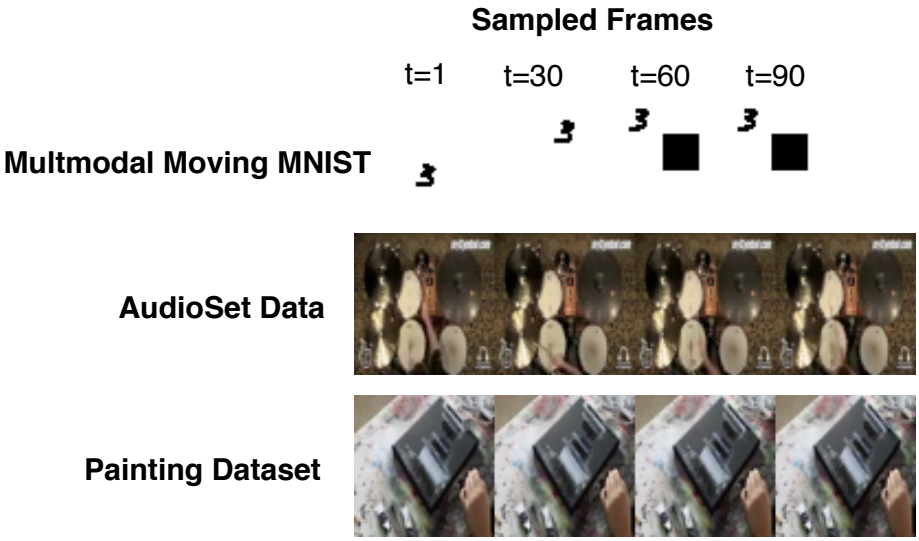


Fig. 2: Sample frames from the three datasets that we used in our experiments.

3 Network Architecture and Training Details

We use an LSTM with 2 layers in the prediction module, the input to which is of 138 dimensions (128 dimensions of features and 10 dimensional z_t). The prior and posterior LSTMs are both single-layered. All LSTMs have a hidden state size of 256 dimensions. Each transformer module has one layer and four heads for capturing multi-head self-attention. The discriminator uses an LSTM with a single hidden layer of 256 dimensions, and a frame-history $R = 2$ and a look-ahead window of size $k = 1$. We train the generator and discriminator jointly

Table 2: Standard Deviation Scores of SSIM, PSNR on the test set of M3SO-NB and M3SO Datasets.

Experiments with M3SO-NB with 5 seen frames							
Method	**Type**	**SSIM**			**PSNR**		
		Frame 6	Frame 15	Frame 25	Frame 6	Frame 15	Frame 25
Our Method	Multimodal	0.0011	0.0051	0.0037	0.157	0.197	0.142
Multiple Frames - [8]	Multimodal	0.0008	0.0016	0.0006	0.032	0.100	0.112
Vougioukas *et al.* [8]	Multimodal	0.0169	0.0175	0.0169	0.069	0.120	0.121
Denton and Fergus [2]	Unimodal - V	0.0009	0.0021	0.0014	0.090	0.064	0.057
Audio Only	Unimodal - A	0.0045	0.0036	0.0034	0.120	0.131	0.102
Experiments on M3SO with 30 seen frames (Block is introduced in the 42^{nd} frame)							
		Frame 31	Frame 42	Frame 70	Frame 31	Frame 42	Frame 70
Our Method	Multimodal	0.0050	0.0064	0.0064	0.178	0.127	0.144
Multiple Frames - [8]	Multimodal	0.0011	0.0029	0.0032	0.142	0.038	0.027
Vougioukas *et al.* [8]	Multimodal	0.0008	0.0024	0.0014	0.076	0.002	0.018
Denton and Fergus [2]	Unimodal - V	0.0106	0.0226	0.0193	0.417	0.182	0.147
Audio Only	Unimodal - A	0.0123	0.0178	0.0179	0.307	0.306	0.307

with an initial learning rate of 0.002 for both, using the ADAM [5] optimizer. During inference, we sample 100 futures per time step, and use the one that maximally matches the ground-truth for evaluating our method. We use the same evaluation for all baseline methods that can generate multiple plausible futures. The weighting term on the KL-loss, β, and the weight on the discriminator loss, γ, were both set to 0.0001 for all datasets. However, γ was increased by a factor of 10 every 300 training epochs. All hyper-parameters were chosen using cross-validation on the validation category of every dataset.

4 Standard Deviation Measures of Model Performance

Tables 2, 3 present the standard deviation of SSIM and PSNR scores on the test set of M3SO-NB, M3SO, AudioSet-Drums, and YouTube Painting datasets. The low standard deviation scores on both SSIM and PSNR, across all datasets, underscore the gains of our method over competing methods. For the mean test set SSIM and PSNR scores on these datasets, we refer the interested reader to Tables 1 and 2 of the main paper.

5 Auxiliary Evaluation of Generated Video Quality

Besides evaluating our method against the baselines using the SSIM and PSNR (see Table 1 and 2 in the main paper), we use some auxiliary measures to further evaluate the quality of synthesis. Table 4 presents the fooling rate of a discriminator trained to distinguish real video clips of length $R + (k - 1)$ (which equals 2 in our case) frames (and their audio) from synthetic ones, on both real world-datasets . We see that our approach outperforms all baselines, attesting to

Table 3: Standard Deviation Scores of SSIM, PSNR on the test set of AudioSet, YouTube Painting Datasets.

Experiments on the AudioSet Dataset [3], with 15 seen frames							
Method	**Type**	**SSIM**			**PSNR**		
		Frame 16	Frame 30	Frame 45	Frame 16	Frame 30	Frame 45
Our Method	Multimodal	0.0092	0.0065	0.0065	0.123	0.519	0.266
Multiple Frames - [8]	Multimodal	0.0168	0.0073	0.0148	0.964	0.231	0.320
Vougioukas *et al.* [8]	Multimodal	0.0162	0.0205	0.2650	0.497	0.402	0.182
Denton and Fergus [2]	Unimodal - V	0.0168	0.0102	0.0084	1.098	0.319	0.054
Hsieh *et al.* [4]	Unimodal - V	0.0050	0.0016	0.0082	0.042	0.051	0.006
Audio Only	Unimodal - A	0.0197	0.0068	0.0072	0.230	0.388	0.177
Experiments on the novel YouTube Painting Dataset, with 15 seen frames							
		Frame 16	Frame 30	Frame 45	Frame 16	Frame 30	Frame 45
Our Method	Multimodal	0.0025	0.0093	0.0146	1.369	0.718	0.250
Multiple Frames - [8]	Multimodal	0.0020	0.0040	0.0050	0.181	0.536	0.879
Vougioukas *et al.* [8]	Multimodal	0.0143	0.0028	0.0150	0.212	0.216	0.449
Denton and Fergus [2]	Unimodal - V	0.0008	0.0193	0.0390	0.431	0.338	0.459
Hsieh *et al.* [4]	Unimodal - V	0.0028	0.0019	0.0030	0.033	0.133	0.150
Audio Only	Unimodal - A	0.0115	0.0069	0.0226	0.353	0.294	0.426

Table 4: Average discriminator fooling rates for different methods on real-world data.

Datasets	**Our Method**	**Denton and Fergus** [2]	**Audio-Only**	**Multiframe** [8]
AudioSet Data	**0.7926**	0.3372	0.5622	0.6095
YouTube Painting Data	**0.6599**	0.4282	0.4687	0.6507

the quality of the frames that are generated by our method. Note that, the discriminator is trained to judge the audio-visual alignment as well as the quality of frames. Thus, a higher discriminator fooling rate implies that the respective model generates more real looking frames, realistic dynamics, and better alignment with audio. Additionally in Table 5, we present the human preference score for samples of our method versus the competitive method of Multiframe [8]. **The results evince that human annotators prefer samples generated by our method overwhelmingly**. For the M3SO dataset, we also measure the intersection over union of the predicted box location against the ground truth location. Table 6 shows that our method outperforms competing methods by a very significant margin (nearly 30%). A high accuracy for this localization task, such as that of our method, demands good capture of the visual dynamics of the digits, along with synchronization against the tones corresponding to the bouncing of the digits with the obstacle.

Table 5: Human preference score on samples generated by our method vs. [8]

Datasets	**Prefer ours**
M3SO- Ours vs. Multiframe [8]	**88%**
AudioSet- Ours vs. Multiframe [8]	**83%**
YouTube Painting-Ours vs. Multiframe [8]	**92%**

Table 6: mIoU on block localization, evaluated for the final frame of the generated sequences on M3SO with block.

Method	**Localization IoU**
Our Method	**0.5801**
Denton and Fergus [2]	0.2577
Vougioukas *et al.* [8]	0.1289
Audio-Only	0.1030

6 Additional Experimental Details

6.1 Ablative Analysis

Figures 3(a) and 3(b) show the performance variations of different ablated variants of our model. Both plots highlight the gains obtained by using: (i) transformer encoder networks [7] to encode the input audio and visual modalities in the stochasticity module; and (ii) the multimodal discrimator network. The transformer encoders help to emphasize the salient components of the features, while attenuating the others by leveraging self-attention. Using the multimodal discriminator discriminator, on the other hand, ensures that the synthesized video clips are realistic and well-aligned with the audio.

Table 7 contrasts our full model against an ablated variant of our model, which does not have the posterior network or the multimodal discriminator (Ours - Only L_2). This variant is trained with only the reconstruction loss (L_2 loss of Equation 5 of the main paper) and serves to disentangle its effect from the other ones in the final objective (Equation 7 in the paper). As is evident from the results, merely training with the reconstruction loss leads to sub-optimal performance.

6.2 Sensitivity of Hyperparameters

Figures 4(a) and 4(b) show the empirical sensitivity analysis of how performance varies with various choices of hyper-parameters β and γ, respectively on the M3SO-NB dataset. As can be seen from these plots, our model attains its peak performance when both these parameters are set to 0.0001.

6.3 Evaluating Diversity of the Prediction Network

Figures 5(a) and 5(b) quantify the diversity in our model's generations. The plots reveal that the more the number of candidate frames (number of futures)

Table 7: SSIM scores for our full model vis-á-vis different variants of our model on M3SO-NB and YouTube Painting Datasets. Highest scores are in **bold**.

Method	**Type**	**M3SO-NB**			**YouTube Painting**		
		Frame 6	Frame 15	Frame 25	Frame 16	Frame 30	Frame 45
Our Method	Multimodal	**0.9575**	**0.8943**	**0.8697**	**0.9716**	0.9291	**0.9110**
Ours - Only L_2	Multimodal	0.9543	0.7624	0.5610	0.9524	0.9166	0.8986
Teacher - Forcing	Multimodal	0.9412	0.8819	0.8519	0.9695	**0.9293**	0.9109
AV Mismatch	Multimodal	0.9428	0.8569	0.8234	0.9496	0.8784	0.8520

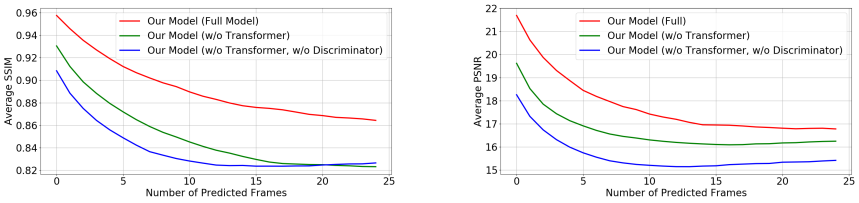
that are sampled at every time step, during inference, the better is the approximation to the ground truth video. However, sampling more frames comes at the cost of computational complexity. Our experiments show that sampling 100 candidate frames at every time step was a good trade-off. We see the pattern of diverse sample generation, carry over to the real world datasets too, as observed in Figures 6(a) and 6(b). Both figures suggest that as the number of candidate futures go up, the diversity of the generated samples increase which is why the average pairwise SSIM between the generated samples tends to go down.

6.4 Performance on a Single Generated Video Sample

Figures 7(a) and 7(b) show the performance of a prototypical sample from the M3SO dataset, vis-á-vis competing state-of-the-art methods. Both these plots show that our approach, outperforms other methods by a significant margin. A closer look into these plots reveal more interesting details. For instance, we notice a sudden dip in performance of all methods at the 12-th frame (frame index number 11). This is due to the sudden appearance of the block at this frame, whose location is not known in advance. However, our model's ability to perceive both audio and visual modalities, allows it to improve its performance as the digit interacts with the updated environment more and more. For instance, a collision with the block is indicated by a sound of a certain tone, which helps to localize the block with respect to the position of the digit. Such useful localization cues are absent in unimodal (vision only) approaches, resulting in poor performance. On the other hand, the approach of Multiframe - Vougioukas *et al.* [8], though multimodal, does not synchronize the audio and visual modalities, and thus fails.

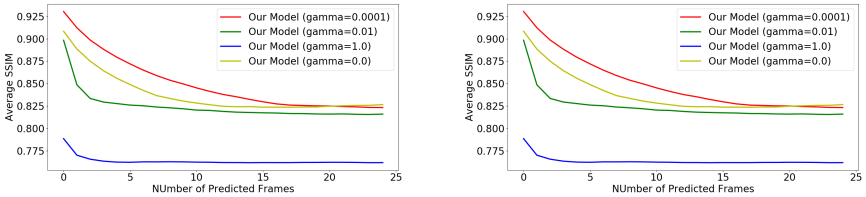
6.5 Training with Teacher Forcing

We considered the impact of using *Teacher-Forcing* for training our model, since such a strategy has shown promise for deterministic sequence-to-sequence models [9]. Here the model is trained with ground-truth frames for the first 100 epochs, but subsequently for every batch, a Bernoulli random variable is sampled to determine if the model is going to be trained with the ground-truth frames (X_t) or by feeding back the synthesized frames ($\hat{X}_t$) as input (as is the case during inference). We observe from the results in Table 7 that for both M3SO-NB



(a) SSIM (b) PSNR

Fig. 3: SSIM (left) and PSNR (right) on the M3SO-NB dataset with ablated variants of our Sound2Sight model.



(a) γ plot (b) β plot

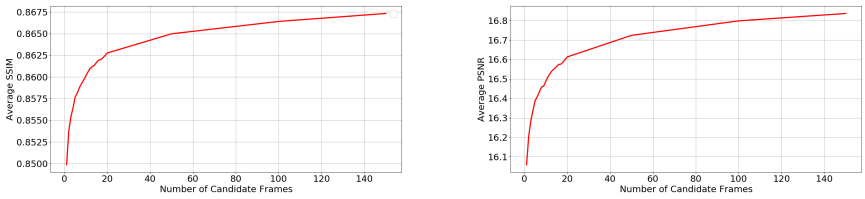
Fig. 4: Hyperparameter Study: Variations in performance on the M3SO-NB dataset against different choices of weighting on the discriminator loss, γ, as measured by SSIM (left) and different choices of weighting on the KL-Divergence loss, β, as measured by SSIM (right)

and for YouTube Painting, the *Teacher-Forcing* variant performs similarly to our original training strategy. We surmise that this is due to the stochastic nature of our model, which permits it to adapt to variations in input data distribution.

6.6 Audio-Video Synchronization

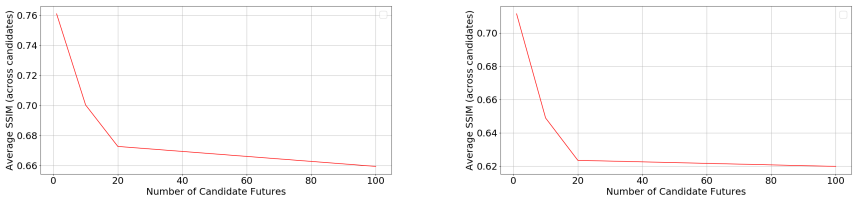
We further validate the ability of our technique to synchronize the audio and visual inputs. A time-evolving SSIM measure of a randomly chosen sample from the test set of M3SO, reveals how our method's predictions can adapt to the stochasticity of the input data (see Figure 8). We refer the interested reader to the attached video to have a better understanding.

Additionally the importance of synchronized audio-visual context is evaluated. In order to do so, we train our model in the standard setup but at inference time we initialize the model with mismatched audio-visual inputs, i.e. the past visual context is not from the same sample from which the audio is drawn. The SSIM scores of the generated frames under this setting is shown in Table 7



(a) SSIM (b) PSNR

Fig. 5: Variation in performance against ground truth video, on the Vanilla MovingMNIST without obstacle with increasing number of candidates (number of futures generated) at every time step, measured by SSIM (left) and PSNR (right).



(a) AudioSet-Drums (b) YouTube-Painting

Fig. 6: Diversity, measured by average SSIM across every pair of future candidate, in the generated frames on AudioSet-Drums (left) and YouTube Paintings (right) with increasing number of candidate futures.

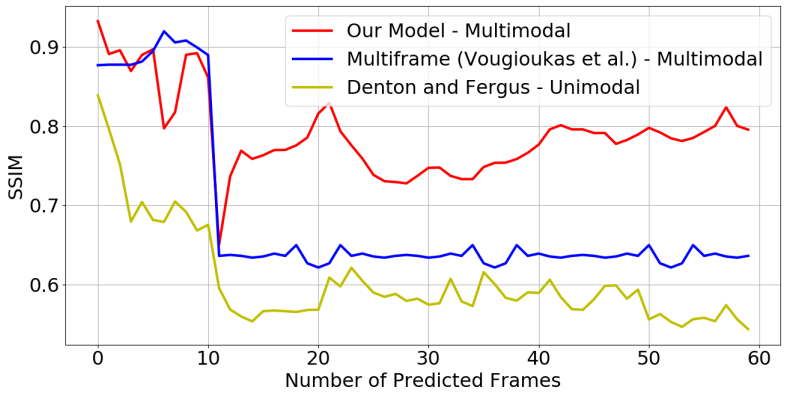
('AV Mismatch'). As is evident from the sub-par performance of this setting, the alignment of the audio and visual inputs is critical for good generation.

7 Qualitative Results

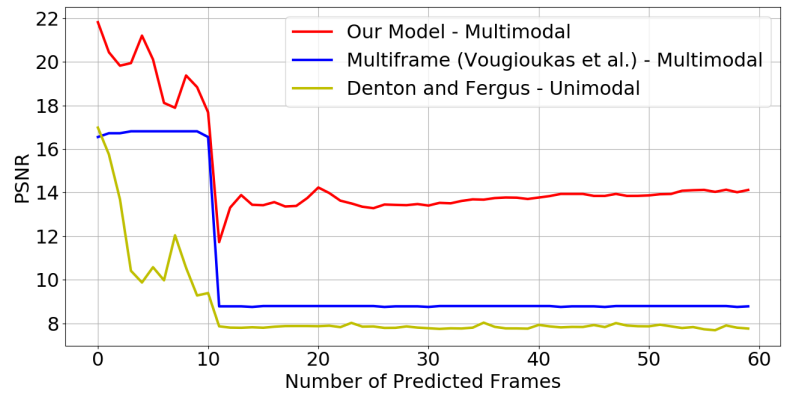
In the following, we present qualitative results of video generation by our method vis-á-vis other state-of-the-art baselines on the Multimodal Moving MNIST dataset (both with and without obstacle) and the real-world YouTube Painting dataset and the AudioSet-Drums [3] datasets. Figures 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19 show visualizations of the frames generated by our method and those by competing baselines on the M3SO-NB dataset. The figures make the case for the superiority of our method in capturing both a digit's appearance and location accurately. We did not observe much of a qualitative difference between the methods of Vougioukas *et al.* [8] and its Multiframe version, across any of the datasets. Hence for brevity, we show only the latter.

Further in the case when the challenges are heightened by introducing an obstacle, we see that our approach demonstrates reasonable performance of localizing both the block and the digit, albeit its appearance is slightly morphed. This stretches beyond the performance of the baseline methods, significantly. In particular, we notice the total disappearance of the digits in each of the baselines. (see Figures 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31). The distinction is more sharply observed, when we see the associated video (attachment in the supplementary materials).

In Figures 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53 we see how our method fares against the state-of-the-art in generating frames from the real world datasets of YouTube Painting and AudioSet-Drums [3], respectively. In all the cases, we see that our method is the closest to the ground-truth and further doesn't introduce artifacts, such as discoloration which some of the baseline methods suffer from. A comparison of the optical flow outputs corroborates this observation. Additionally, the relative crispness of the hand region is suggestive of the fact that our approach is also better at modeling the dynamics of the video. We recommend the interested reader to the accompanying videos to gain a better understanding of the generation performance.



(a) SSIM



(b) PSNR

Fig. 7: Performance on a random sample from the MovingMNIST with obstacle dataset vis-á-vis the most competitive baselines, measured by SSIM (top) and PSNR (bottom).

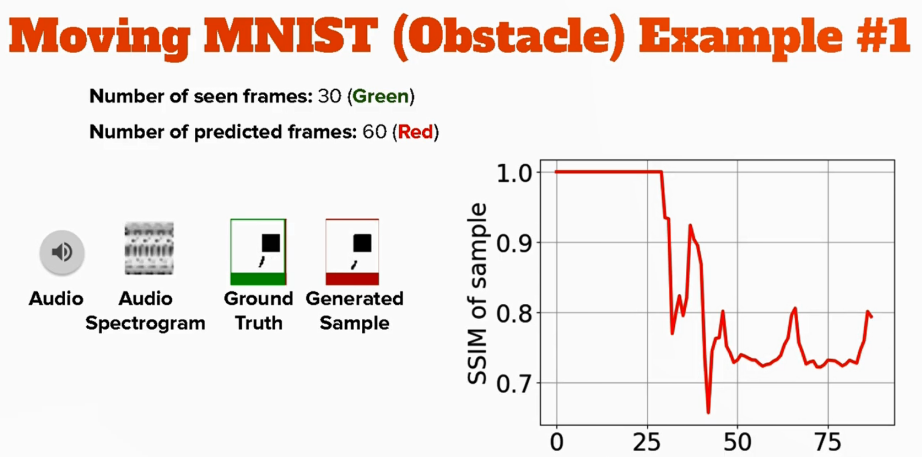


Fig. 8: A screenshot from the video illustrating how the frames predicted correlate with the ground-truth (as measured by SSIM) and the input audio signal. **Please see the attached video clip.**

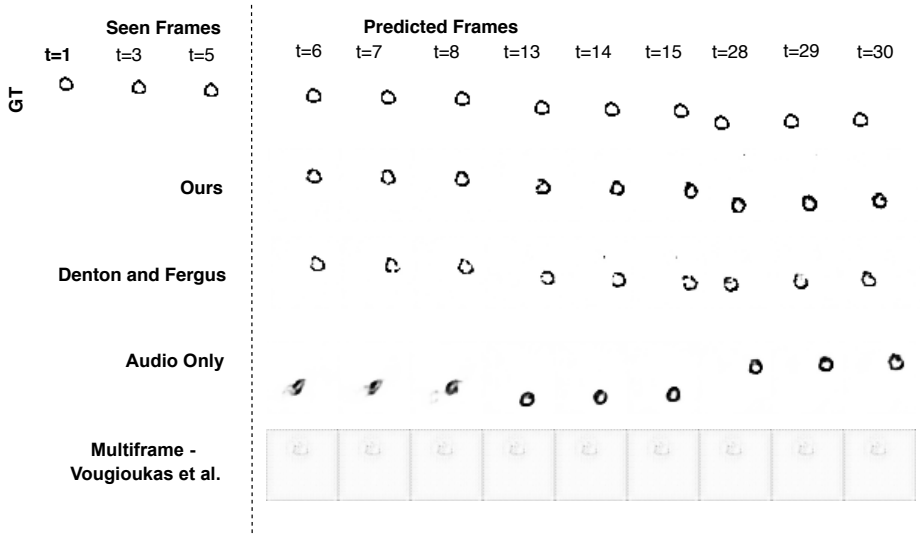


Fig. 9: Sample generations on the M3SO-NB dataset by our method vis-á-vis other baselines.

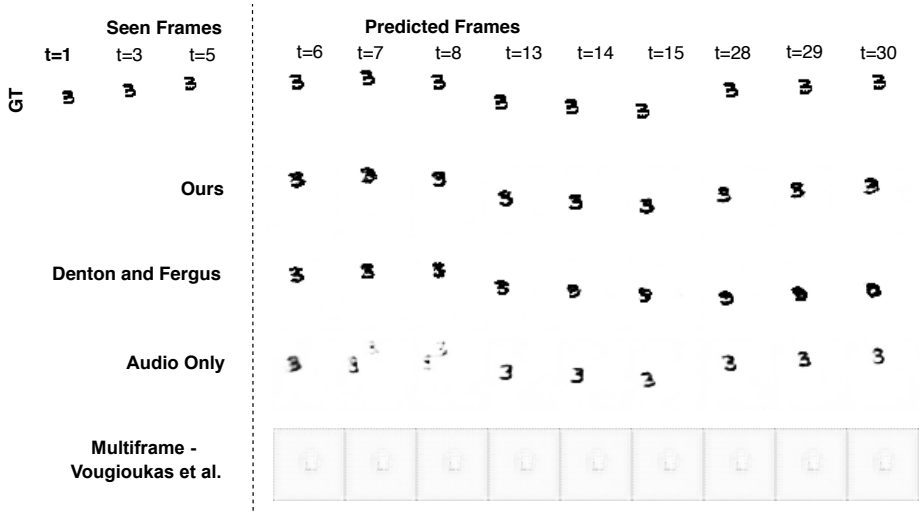


Fig. 10: Sample generations on the M3SO-NB dataset by our method vis-á-vis other baselines.

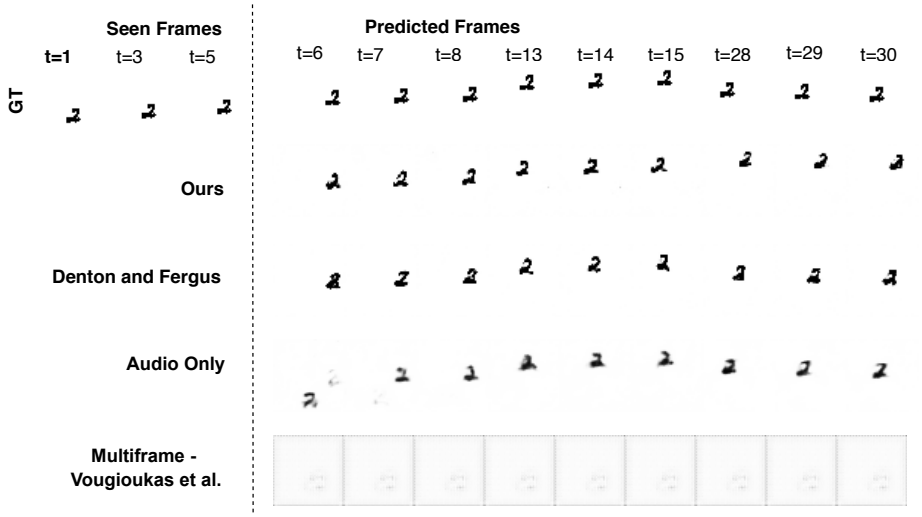


Fig. 11: Sample generations on the M3SO-NB dataset by our method vis-á-vis other baselines.

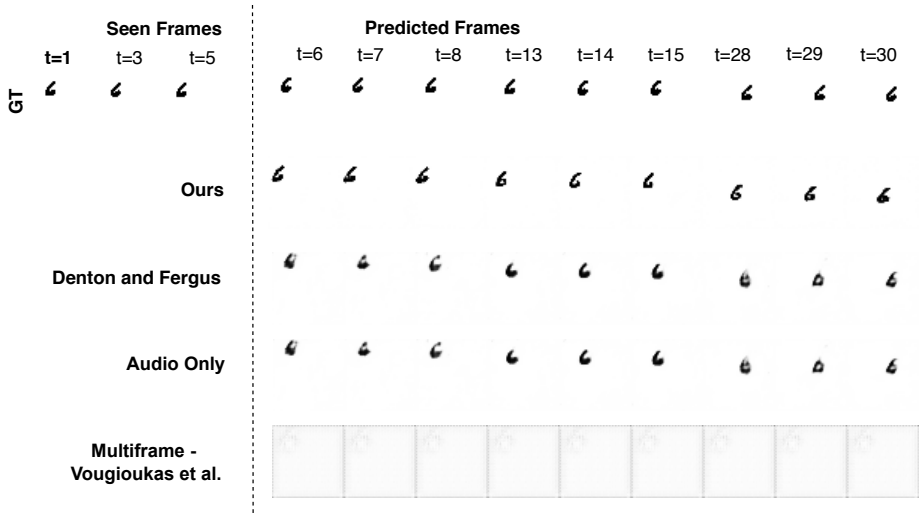


Fig. 12: Sample generations on the M3SO-NB dataset by our method vis-á-vis other baselines.

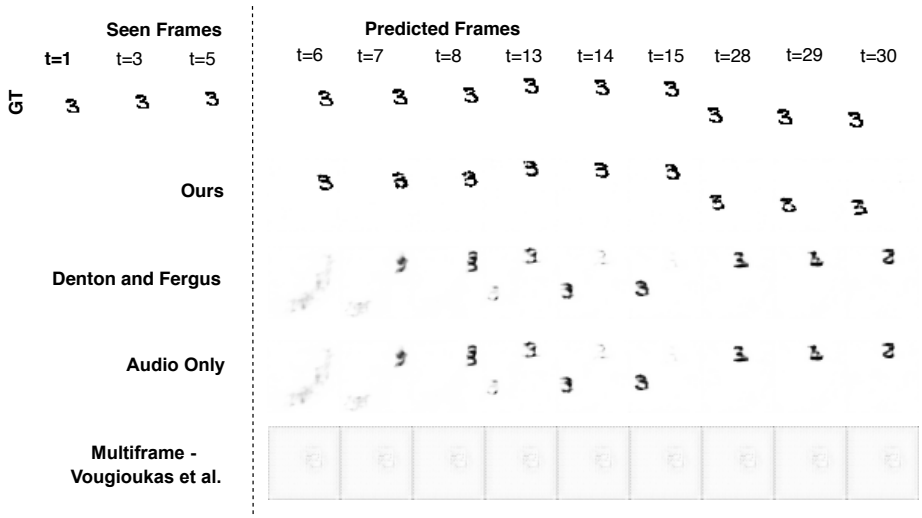


Fig. 13: Sample generations on the M3SO-NB dataset by our method vis-á-vis other baselines.

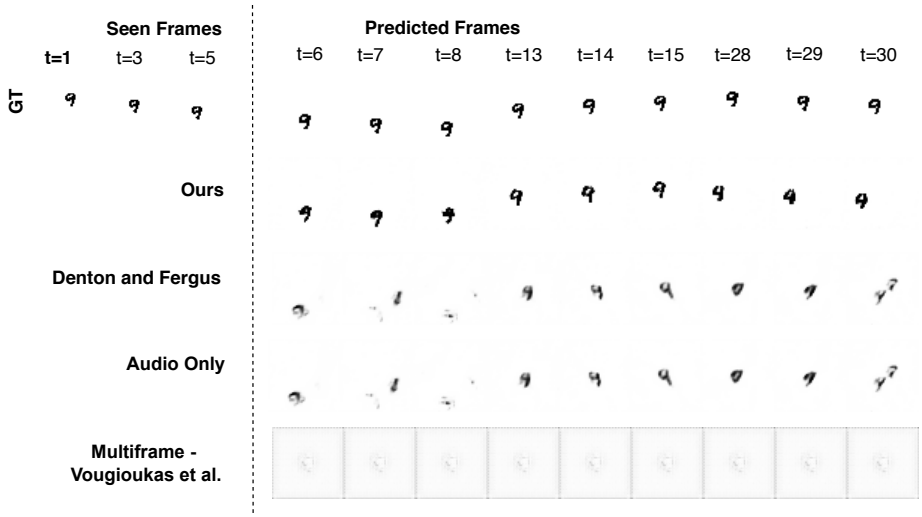


Fig. 14: Sample generations on the M3SO-NB dataset by our method vis-á-vis other baselines.

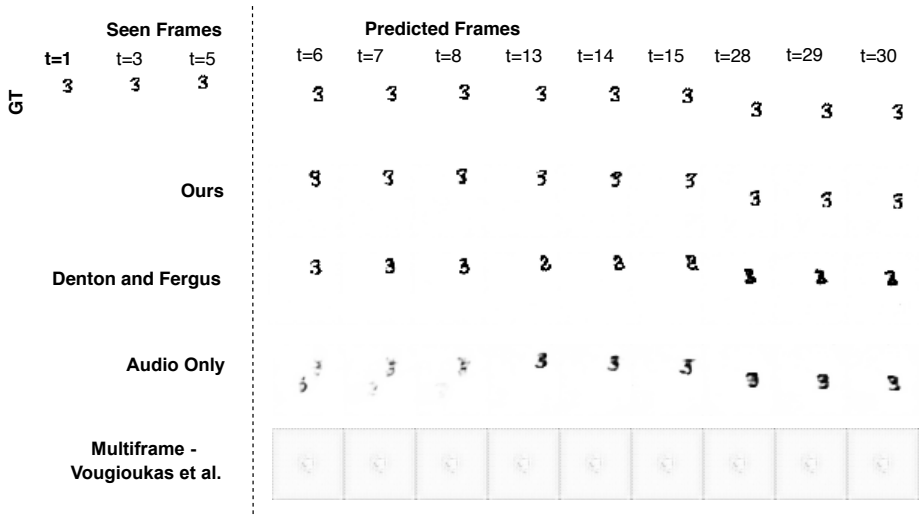


Fig. 15: Sample generations on the M3SO-NB dataset by our method vis-á-vis other baselines.

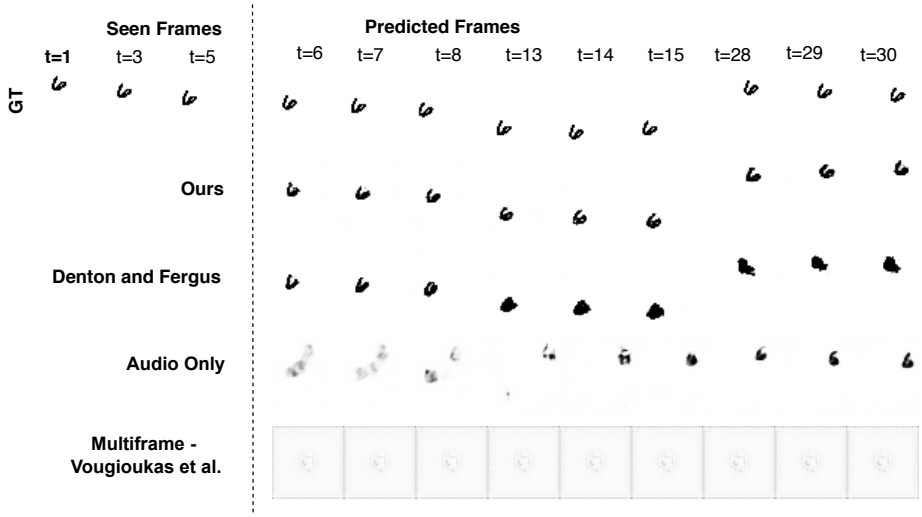


Fig. 16: Sample generations on the M3SO-NB dataset by our method vis-á-vis other baselines.

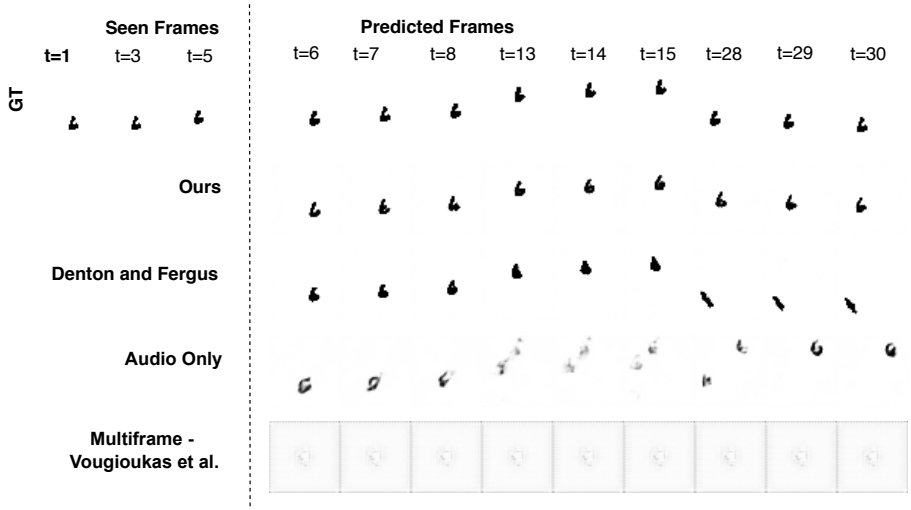


Fig. 17: Sample generations on the M3SO-NB dataset by our method vis-á-vis other baselines.

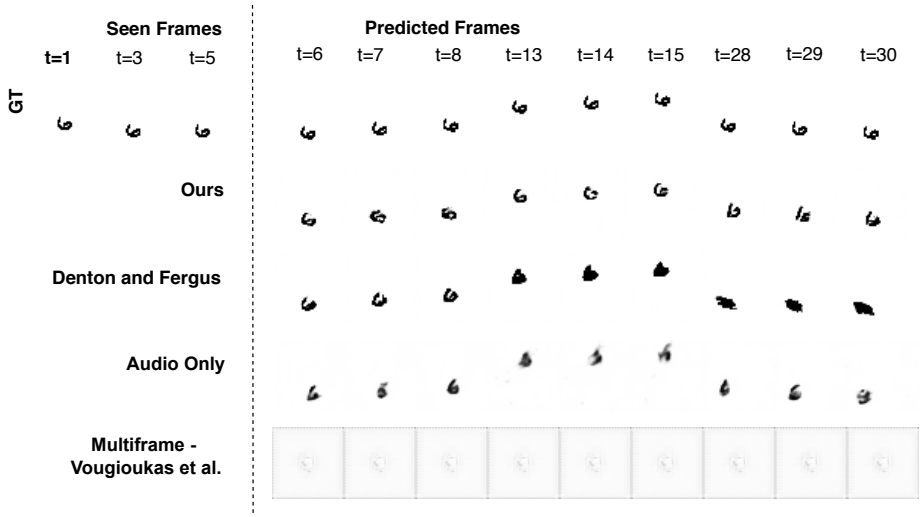


Fig. 18: Sample generations on the M3SO-NB dataset by our method vis-á-vis other baselines.

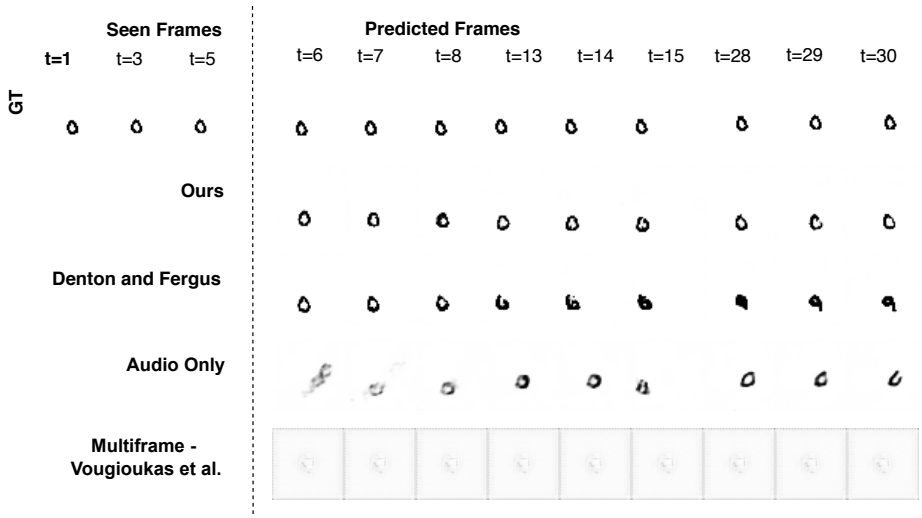


Fig. 19: Sample generations on the M3SO-NB dataset by our method vis-á-vis other baselines.

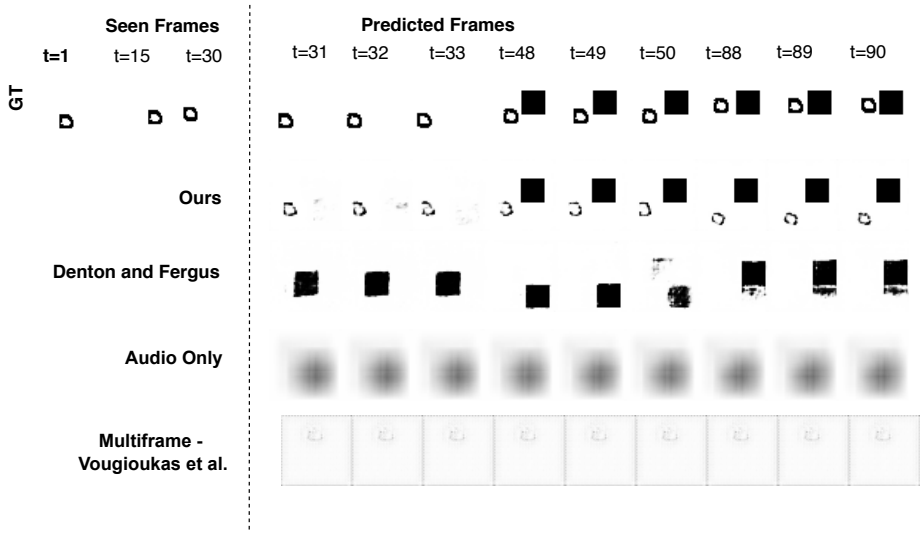


Fig. 20: Sample generations on the MovingMNIST with Surprise Obstacle dataset by our method vis-á-vis other baselines.

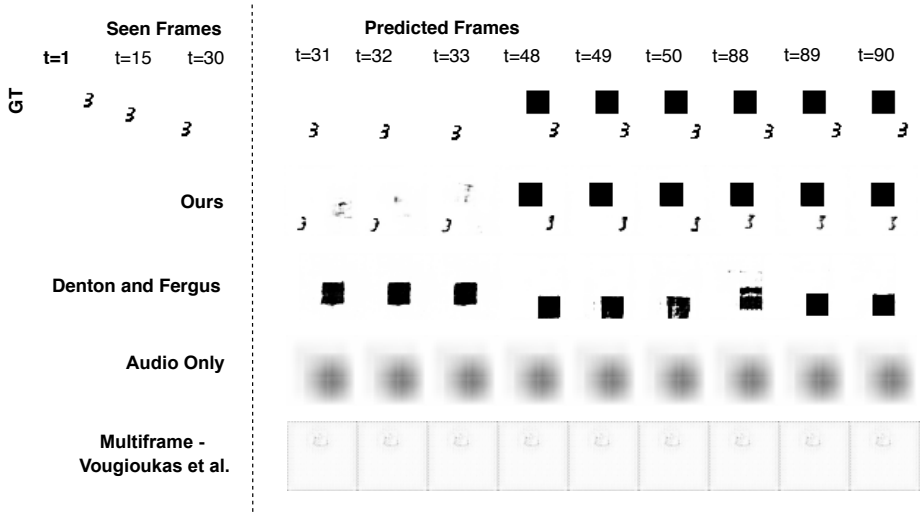


Fig. 21: Sample generations on the MovingMNIST with Surprise Obstacle dataset by our method vis-á-vis other baselines.

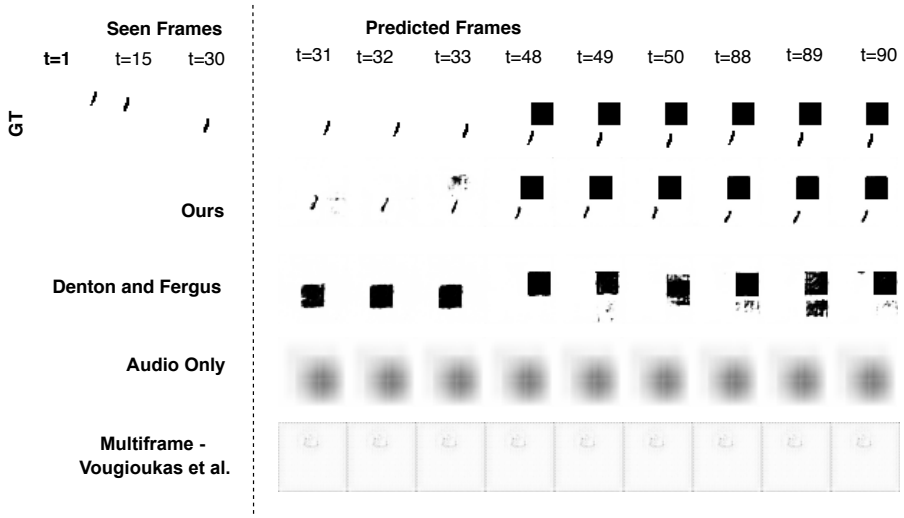


Fig. 22: Sample generations on the MovingMNIST with Surprise Obstacle dataset by our method vis-á-vis other baselines.

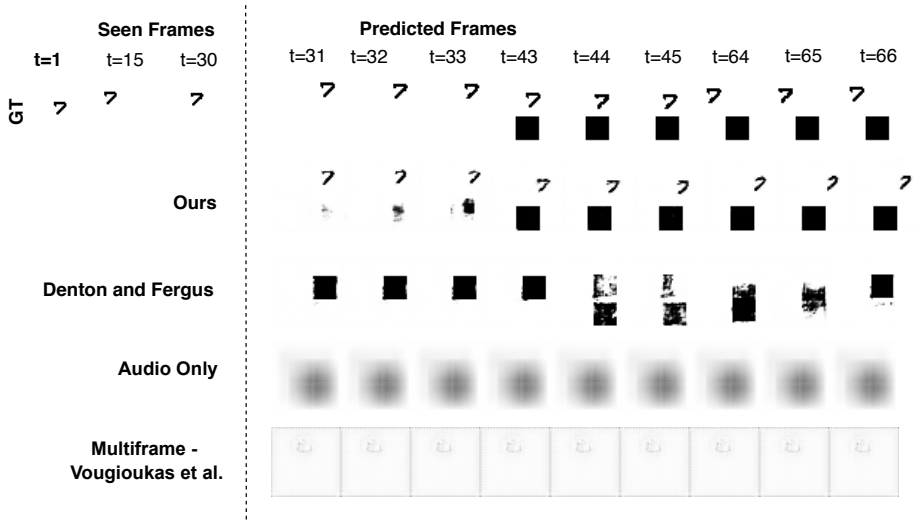


Fig. 23: Sample generations on the MovingMNIST with Surprise Obstacle dataset by our method vis-á-vis other baselines.

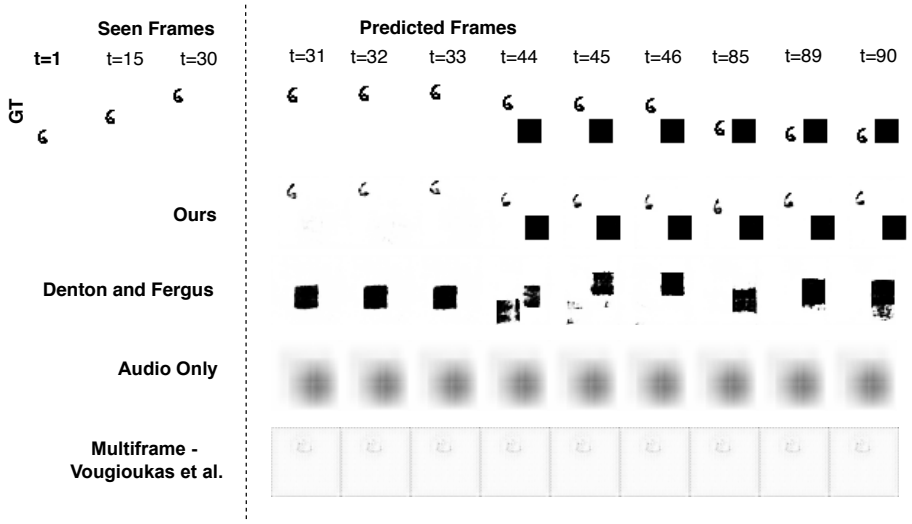


Fig. 24: Sample generations on the MovingMNIST with Surprise Obstacle dataset by our method vis-á-vis other baselines.

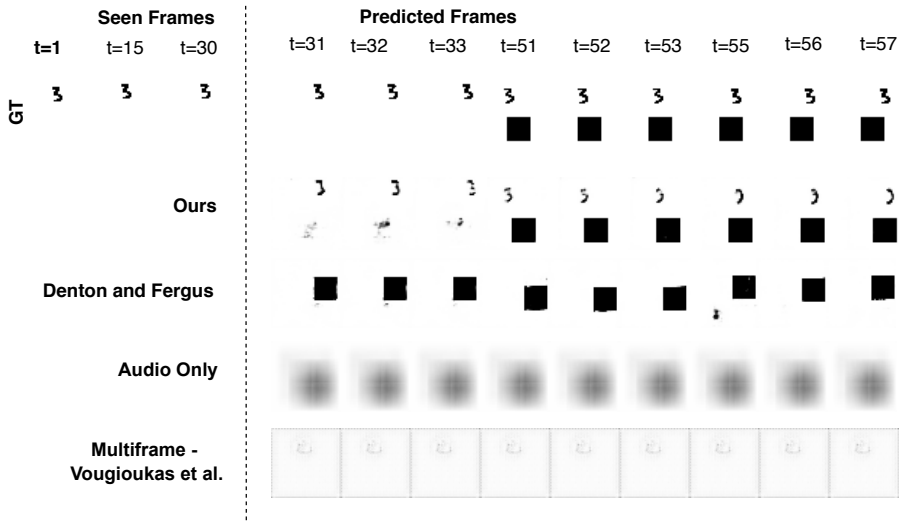


Fig. 25: Sample generations on the MovingMNIST with Surprise Obstacle dataset by our method vis-á-vis other baselines.

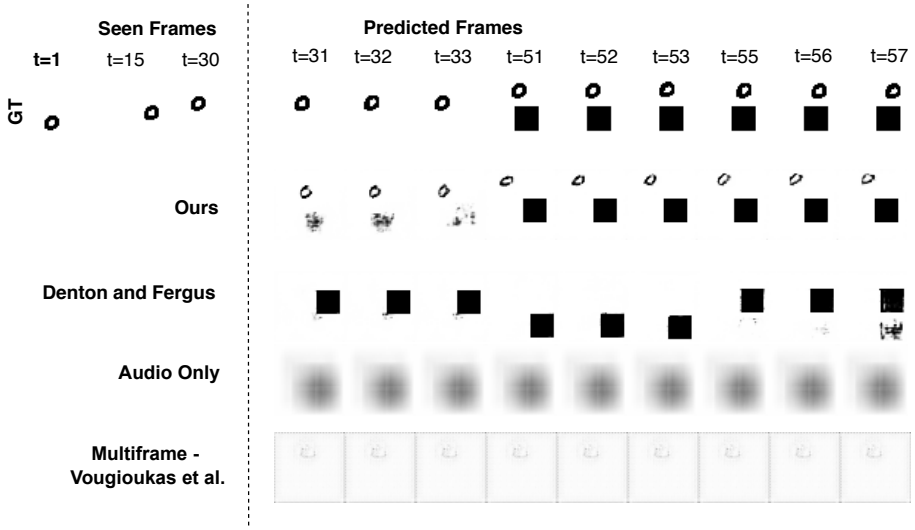


Fig. 26: Sample generations on the MovingMNIST with Surprise Obstacle dataset by our method vis-á-vis other baselines.

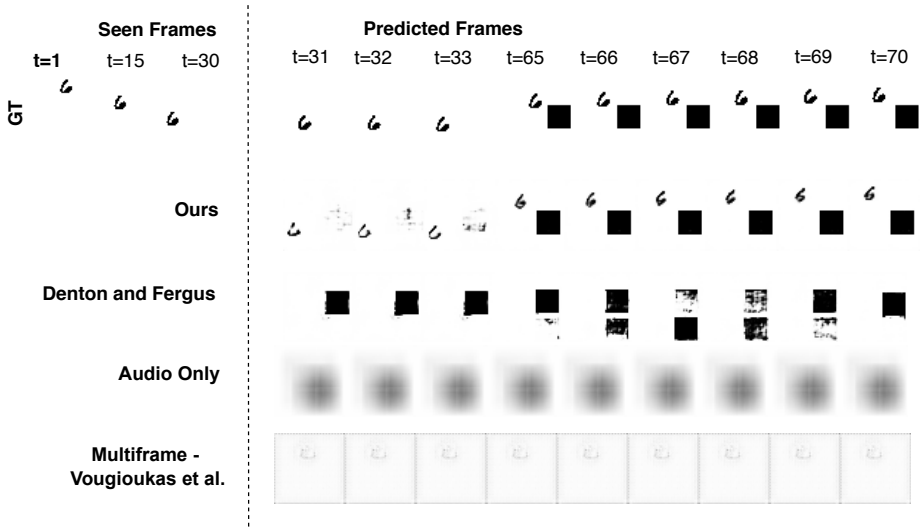


Fig. 27: Sample generations on the MovingMNIST with Surprise Obstacle dataset by our method vis-á-vis other baselines.

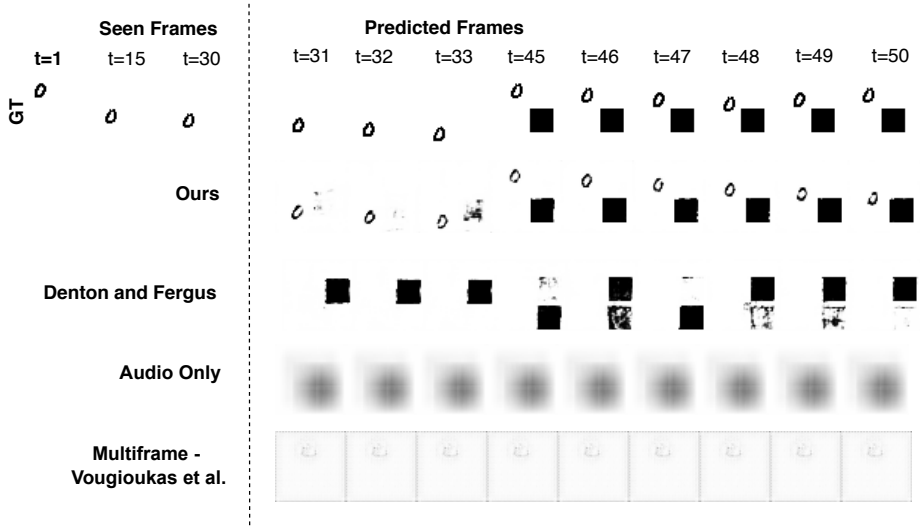


Fig. 28: Sample generations on the MovingMNIST with Surprise Obstacle dataset by our method vis-á-vis other baselines.

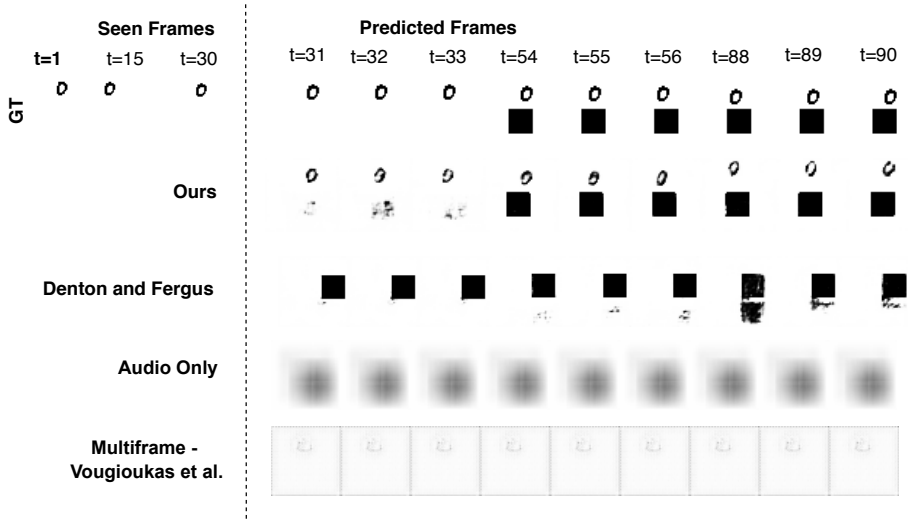


Fig. 29: Sample generations on the MovingMNIST with Surprise Obstacle dataset by our method vis-á-vis other baselines.

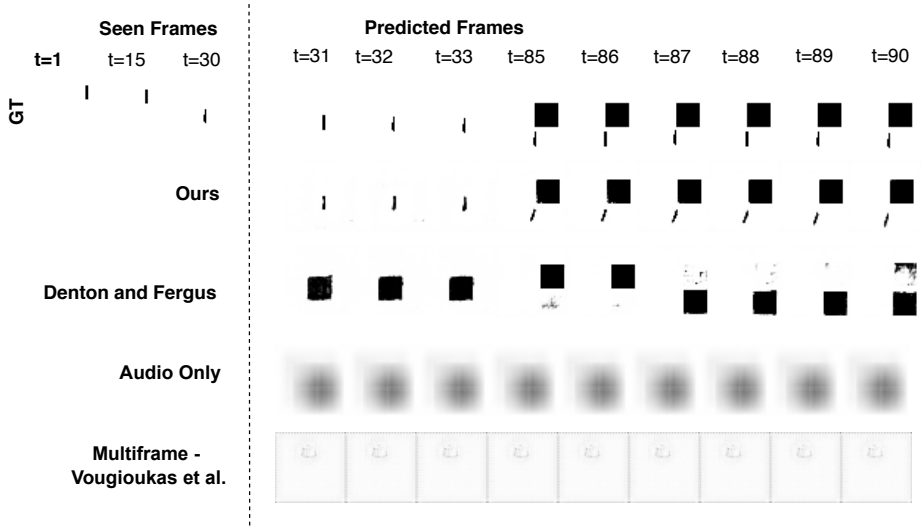


Fig. 30: Sample generations on the MovingMNIST with Surprise Obstacle dataset by our method vis-á-vis other baselines.

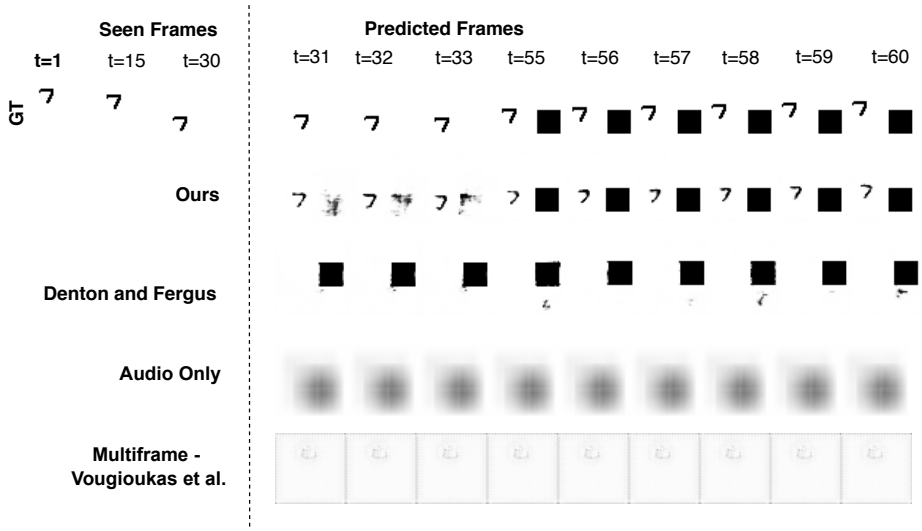


Fig. 31: Sample generations on the MovingMNIST with Surprise Obstacle dataset by our method vis-á-vis other baselines.

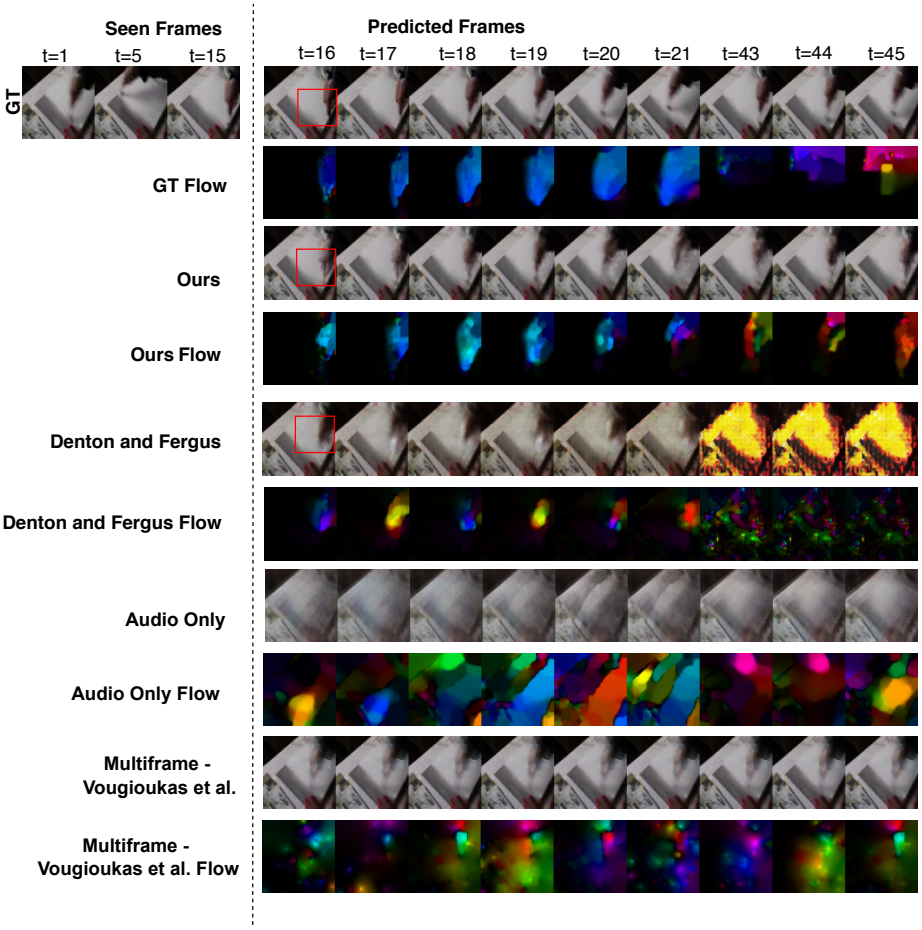


Fig. 32: Sample generations from the YouTube-Painting dataset by our method vis-á-vis other baselines and optical flows across frames. The red squares denotes regions of high motion.

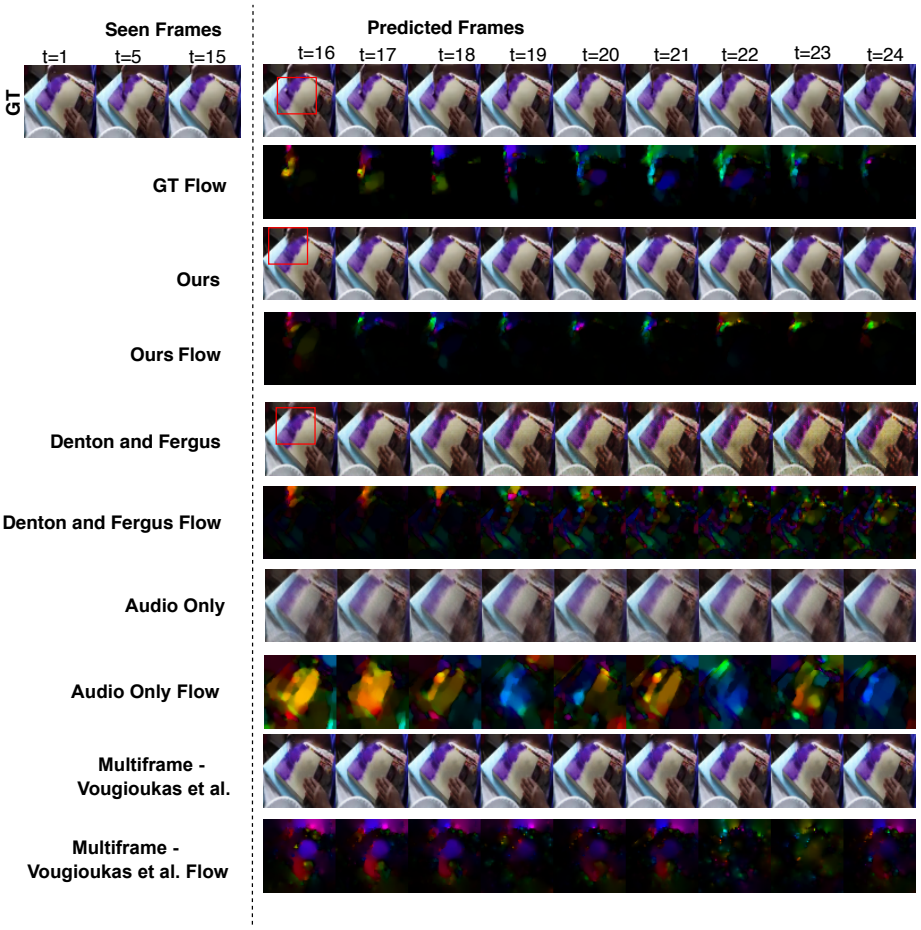


Fig. 33: Sample generations from the YouTube-Painting dataset by our method vis-á-vis other baselines and optical flows across frames. The red squares denotes regions of high motion.

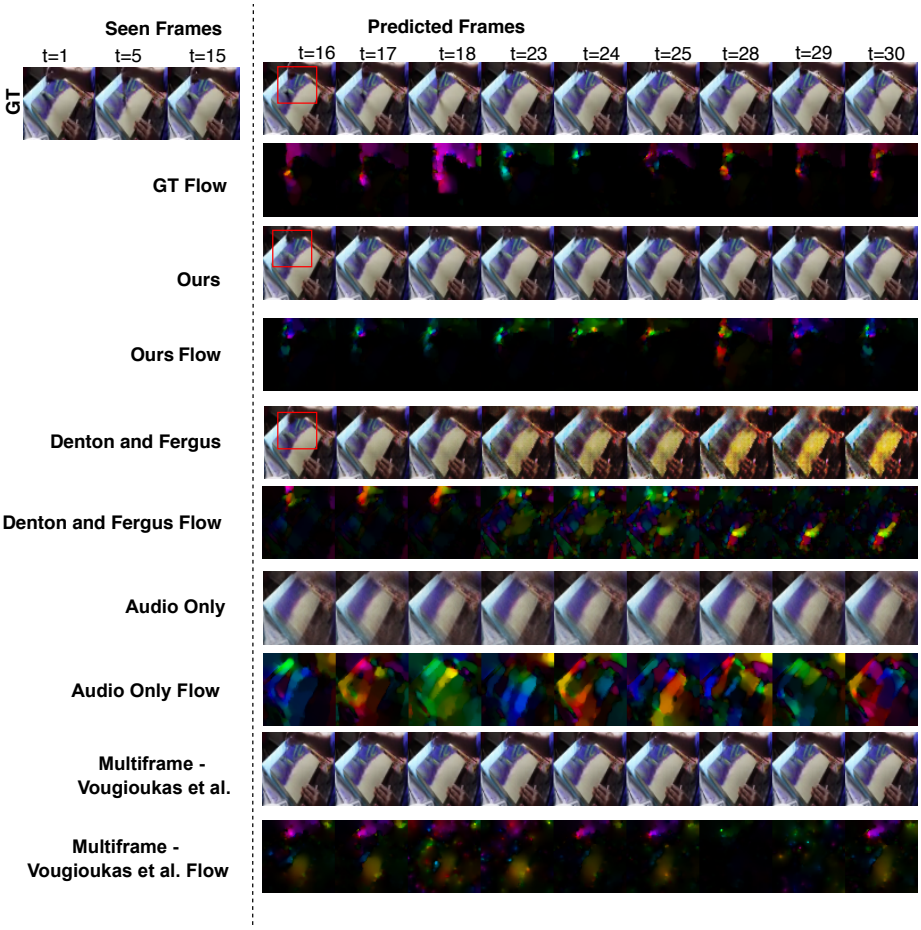


Fig. 34: Sample generations from the YouTube-Painting dataset by our method vis-á-vis other baselines and optical flows across frames. The red squares denotes regions of high motion.

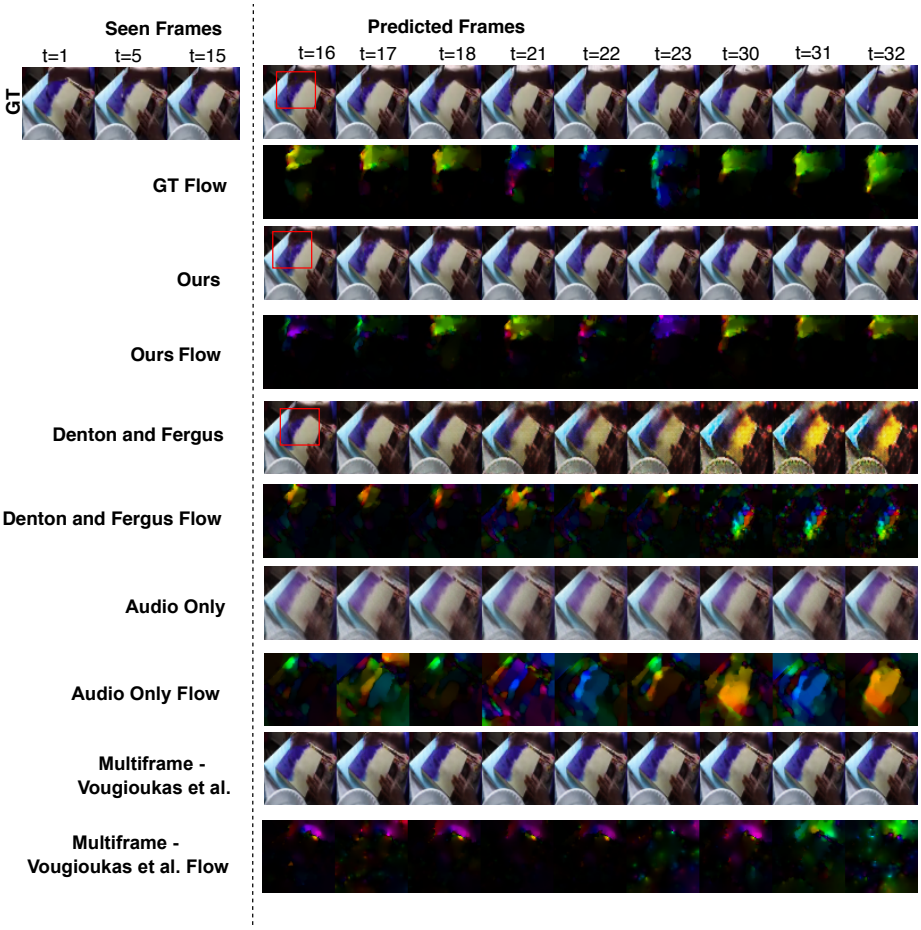


Fig. 35: Sample generations from the YouTube-Painting dataset by our method vis-á-vis other baselines and optical flows across frames. The red squares denotes regions of high motion.

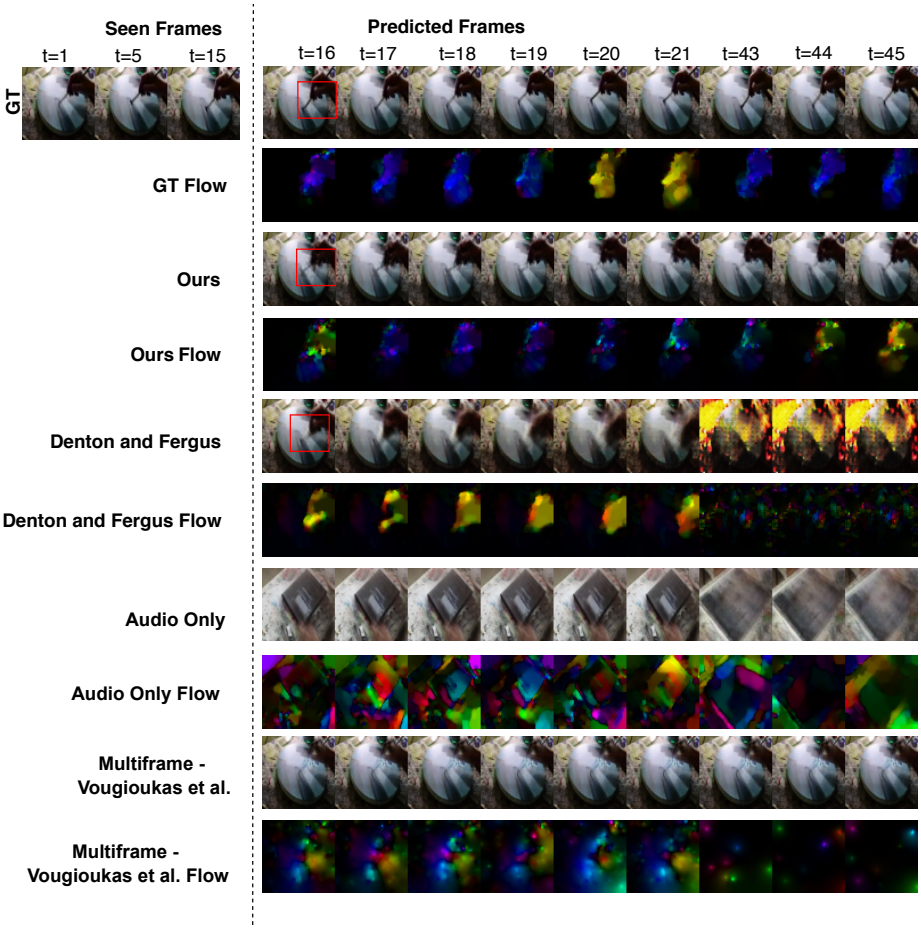


Fig. 36: Sample generations from the YouTube-Painting dataset by our method vis-á-vis other baselines and optical flows across frames. The red squares denotes regions of high motion.

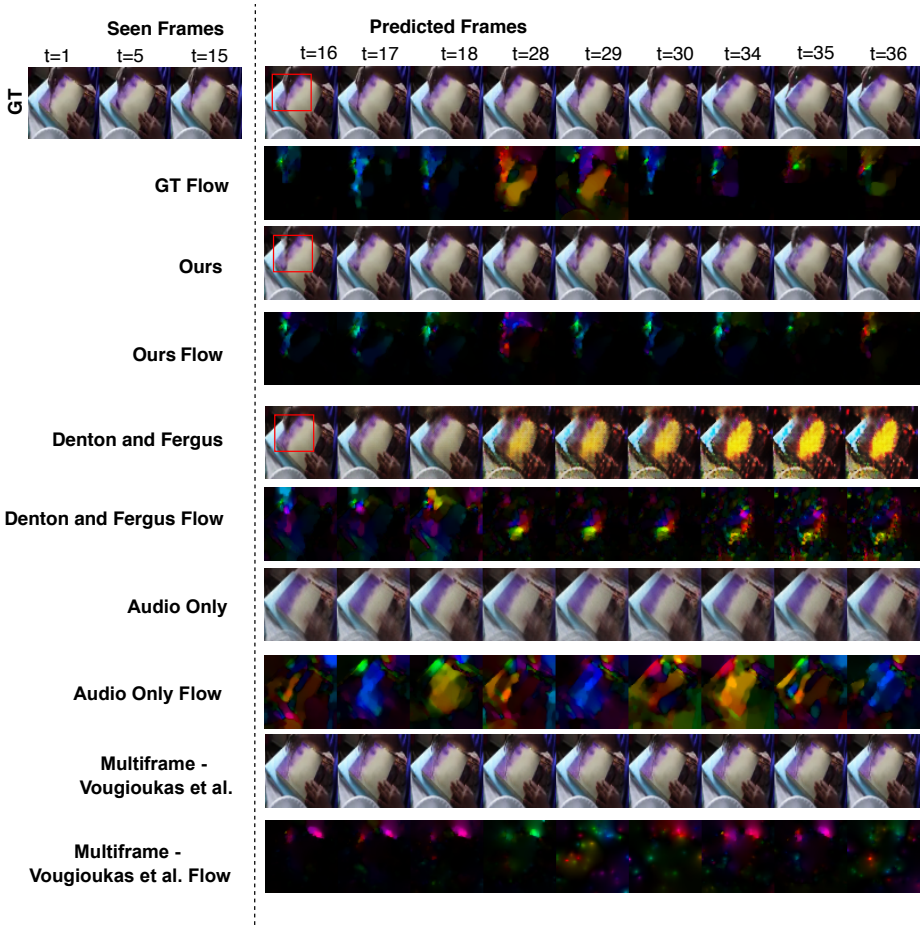


Fig. 37: Sample generations from the YouTube-Painting dataset by our method vis-á-vis other baselines and optical flows across frames. The red squares denotes regions of high motion.

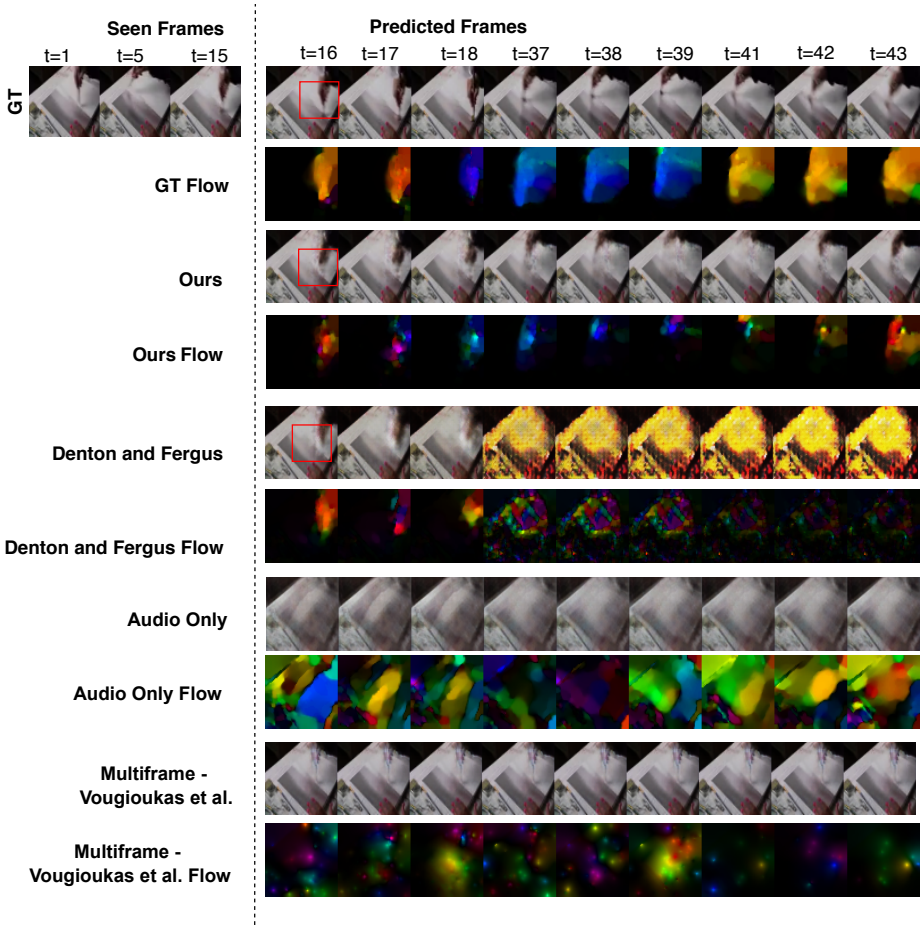


Fig. 38: Sample generations from the YouTube-Painting dataset by our method vis-á-vis other baselines and optical flows across frames. The red squares denotes regions of high motion.

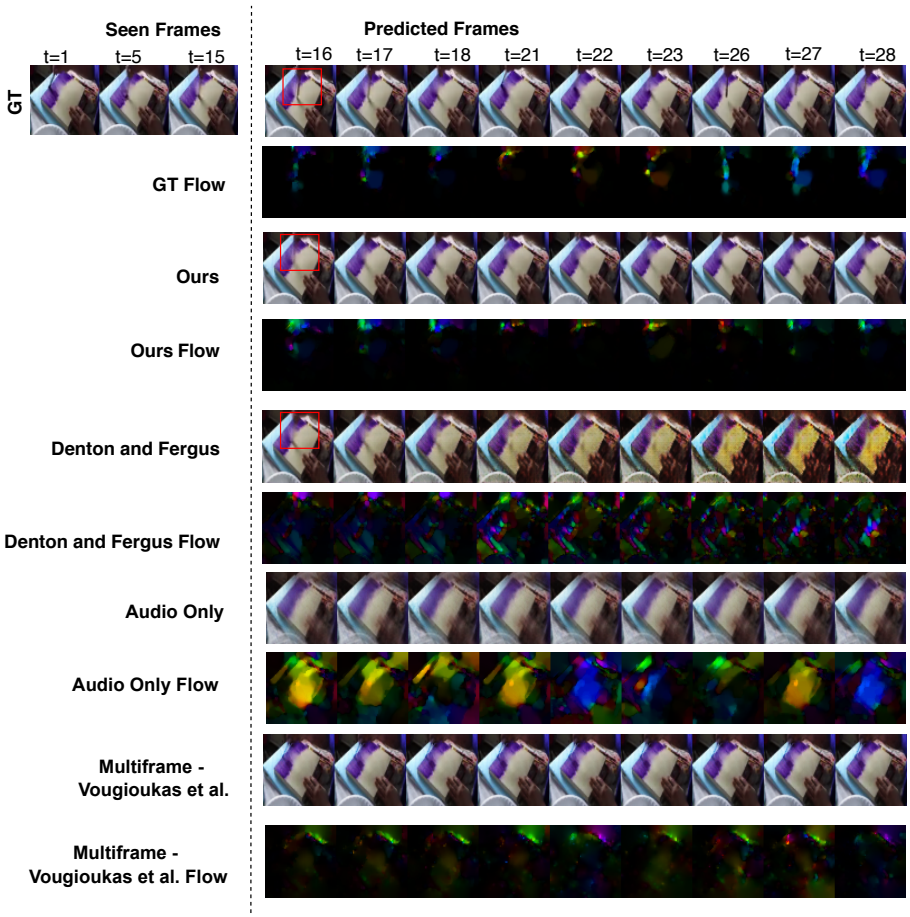


Fig. 39: Sample generations from the YouTube-Painting dataset by our method vis-á-vis other baselines and optical flows across frames. The red squares denotes regions of high motion.

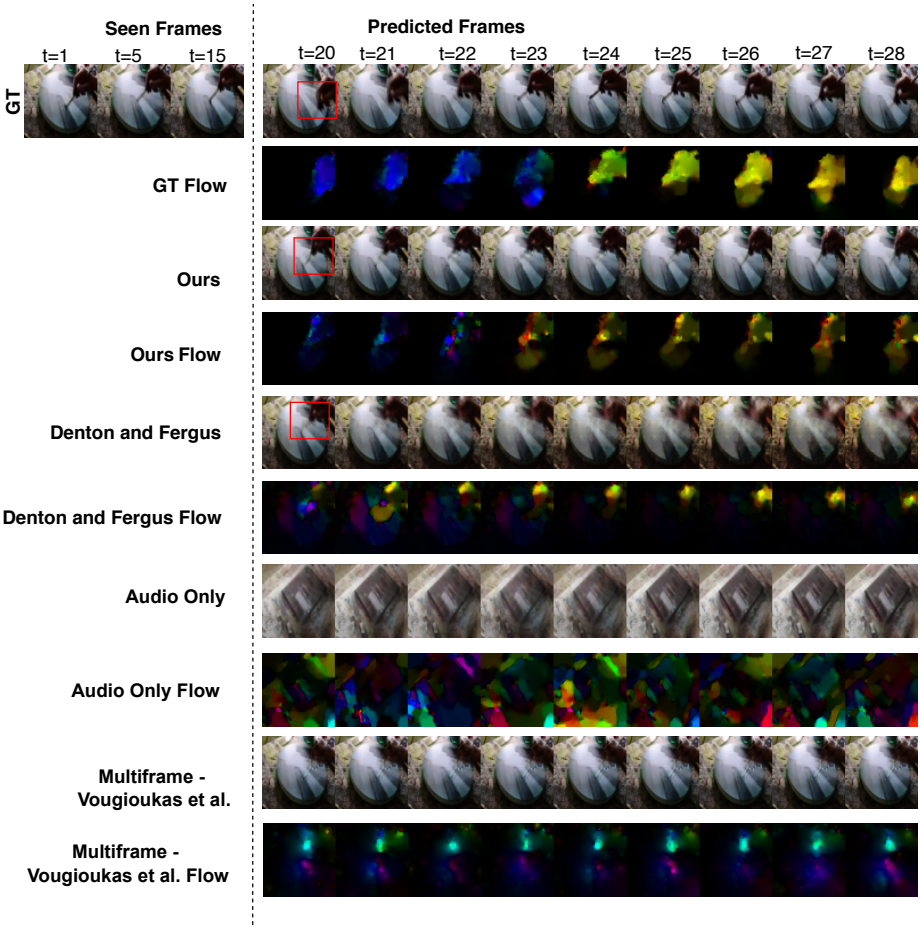


Fig. 40: Sample generations from the YouTube-Painting dataset by our method vis-á-vis other baselines and optical flows across frames. The red squares denotes regions of high motion.

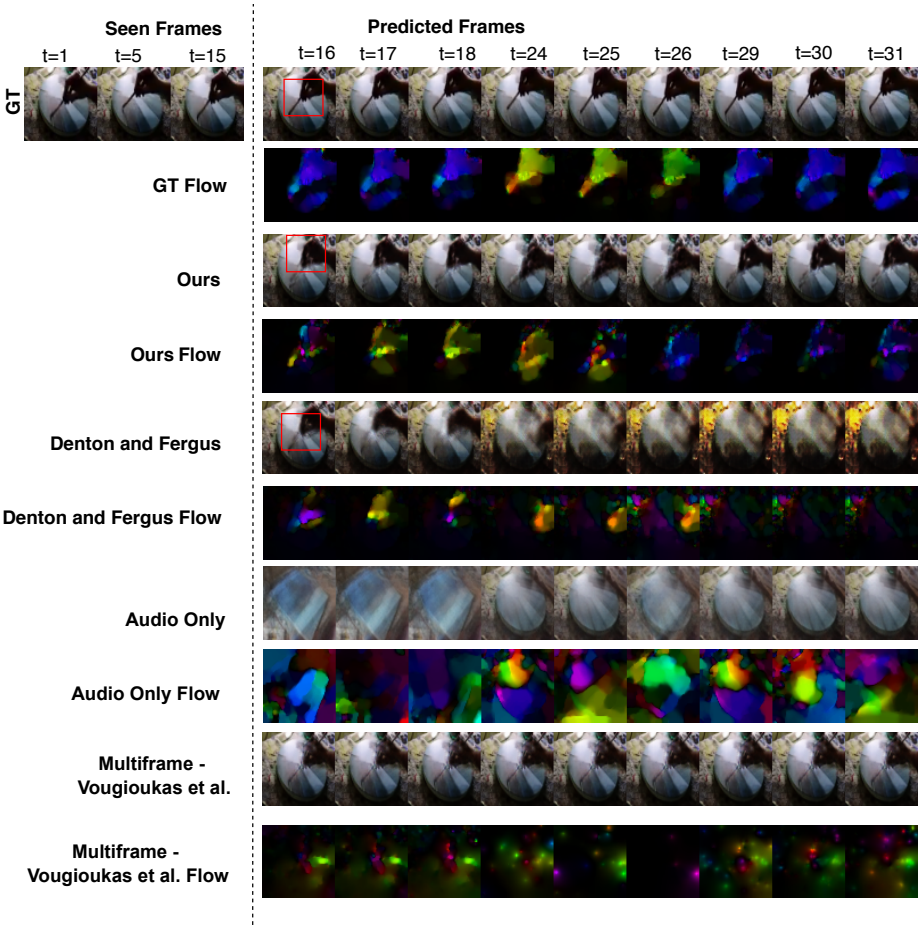


Fig. 41: Sample generations from the YouTube-Painting dataset by our method vis-á-vis other baselines and optical flows across frames. The red squares denotes regions of high motion.

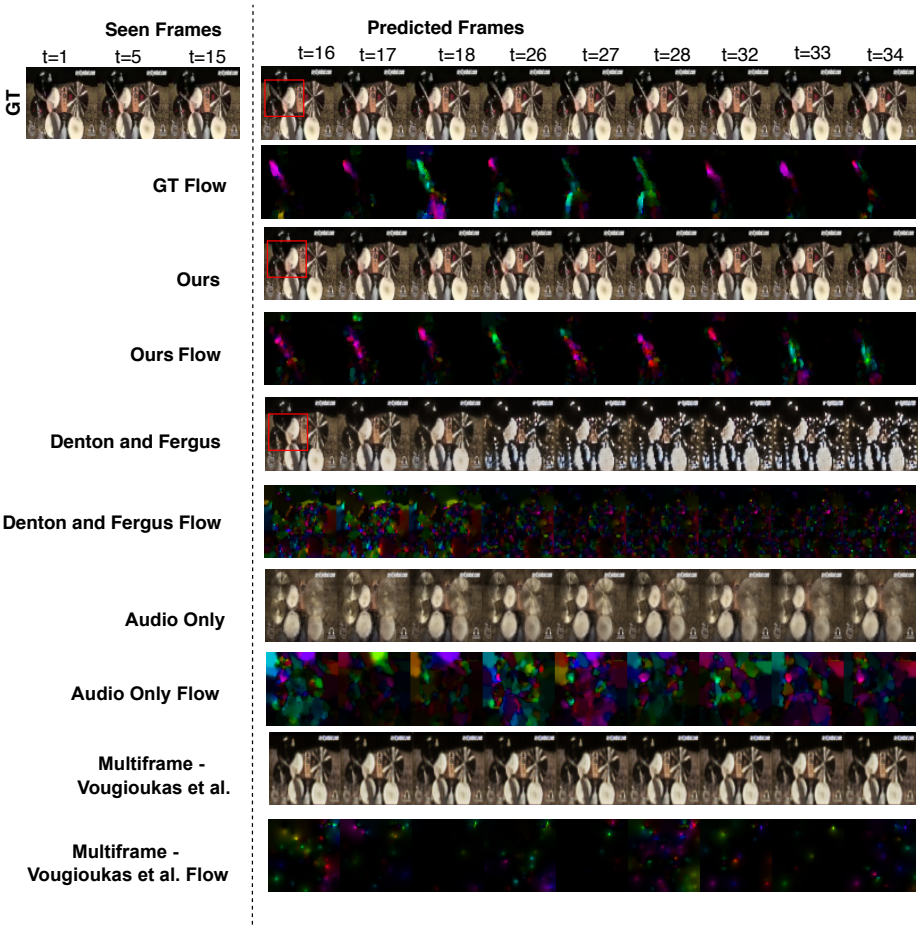


Fig. 42: Sample generations on the AudioSet-Drums dataset by our method vis-á-vis other baselines and optical flows across frames. The red square denotes regions of high motion.

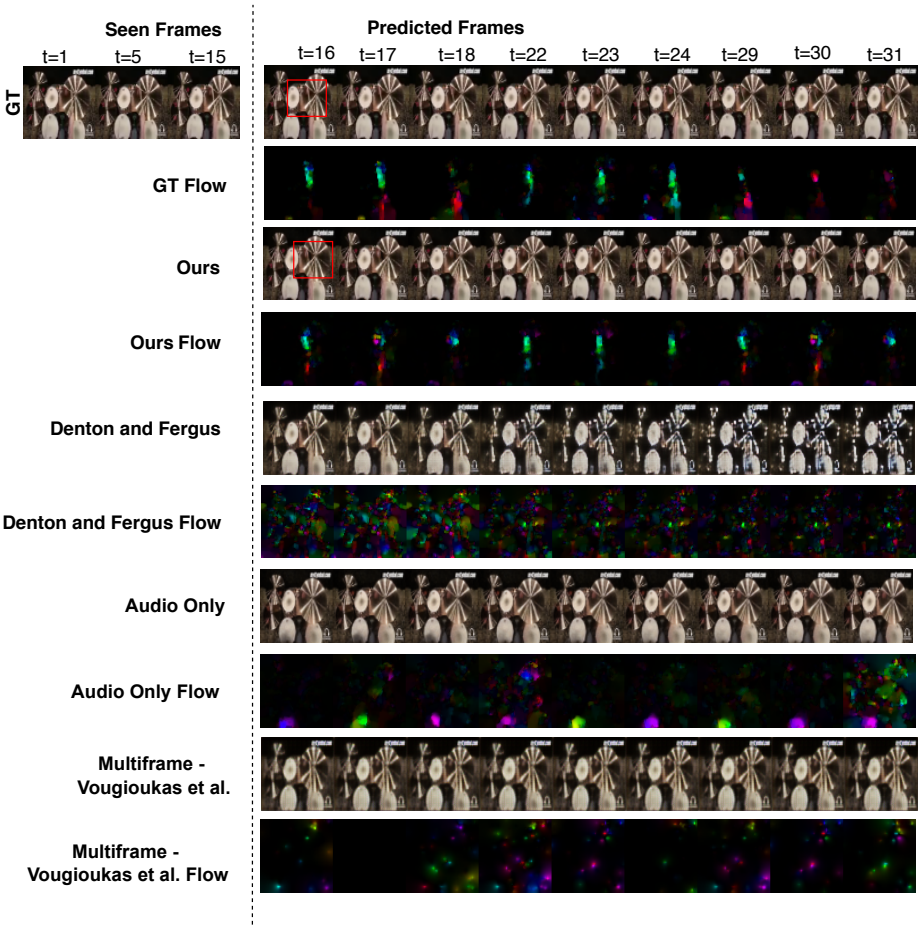


Fig. 43: Sample generations on the AudioSet-Drums dataset by our method vis-á-vis other baselines and optical flows across frames. The red square denotes regions of high motion.

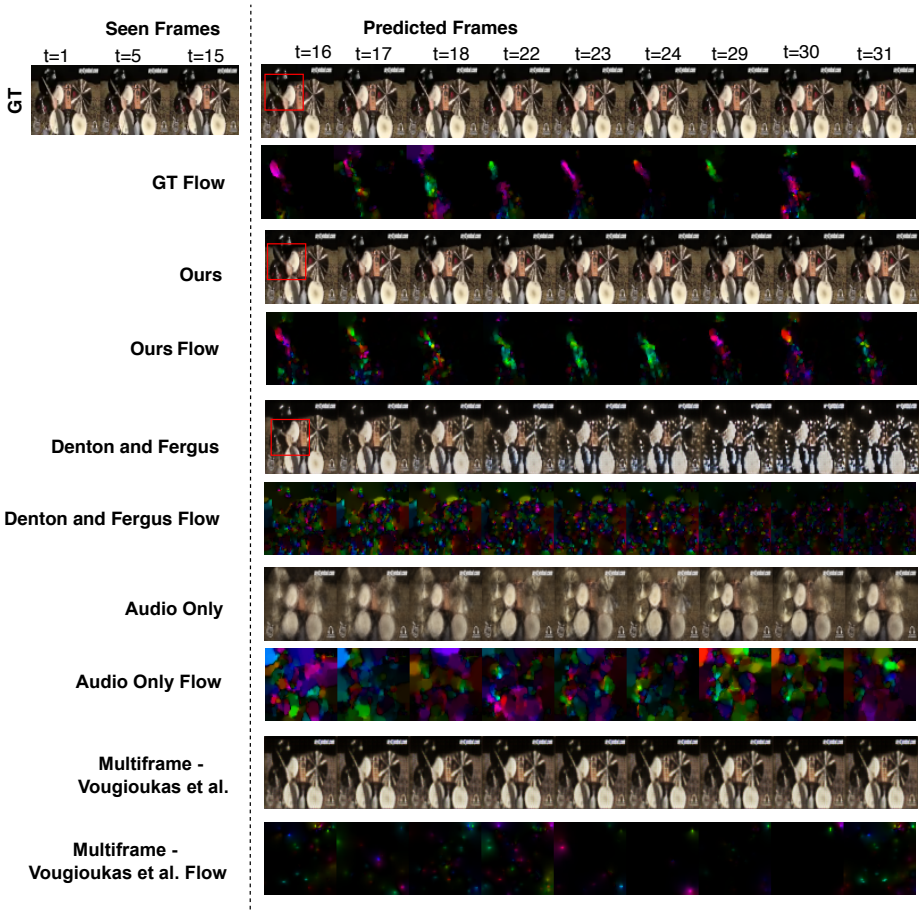


Fig. 44: Sample generations on the AudioSet-Drums dataset by our method vis-á-vis other baselines and optical flows across frames. The red square denotes regions of high motion.

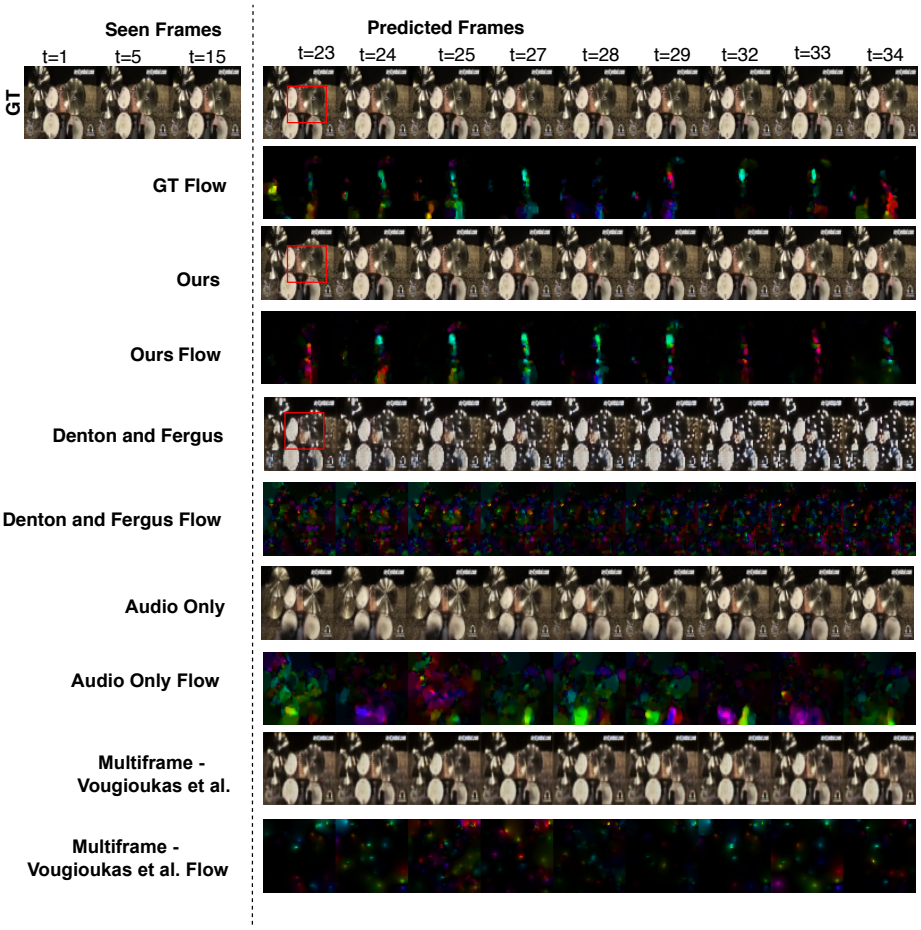


Fig. 45: Sample generations on the AudioSet-Drums dataset by our method vis-á-vis other baselines and optical flows across frames. The red square denotes regions of high motion.

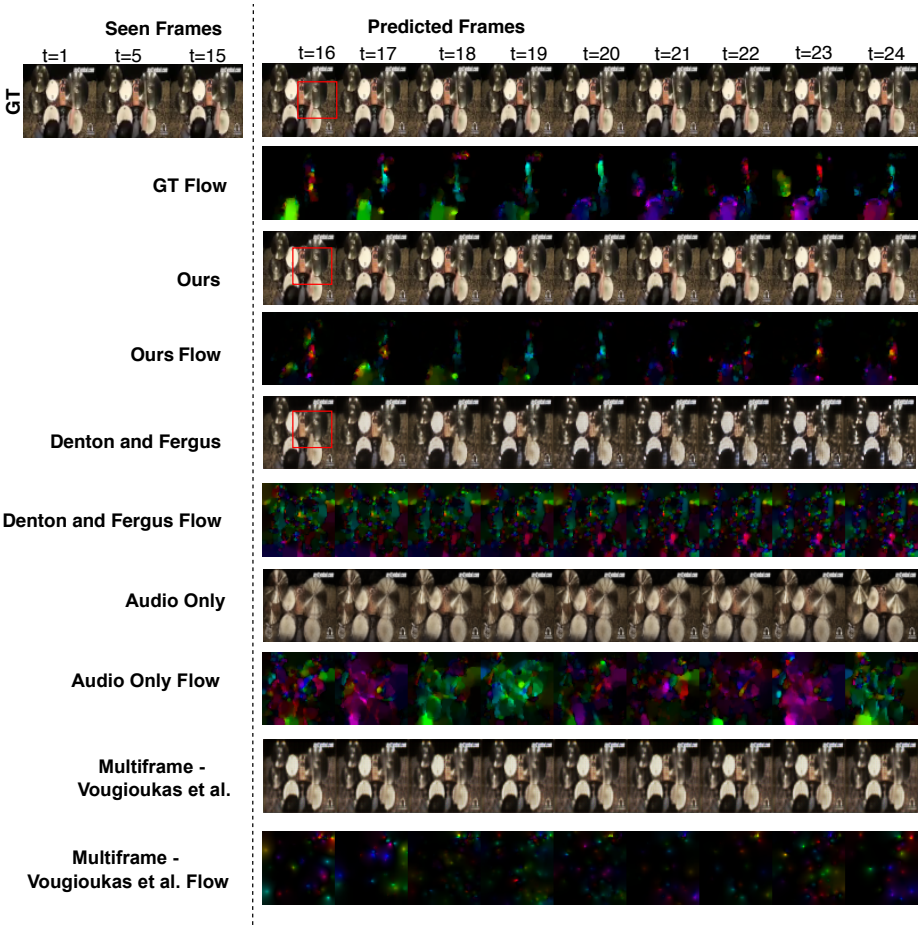


Fig. 46: Sample generations on the AudioSet-Drums dataset by our method vis-á-vis other baselines and optical flows across frames. The red square denotes regions of high motion.

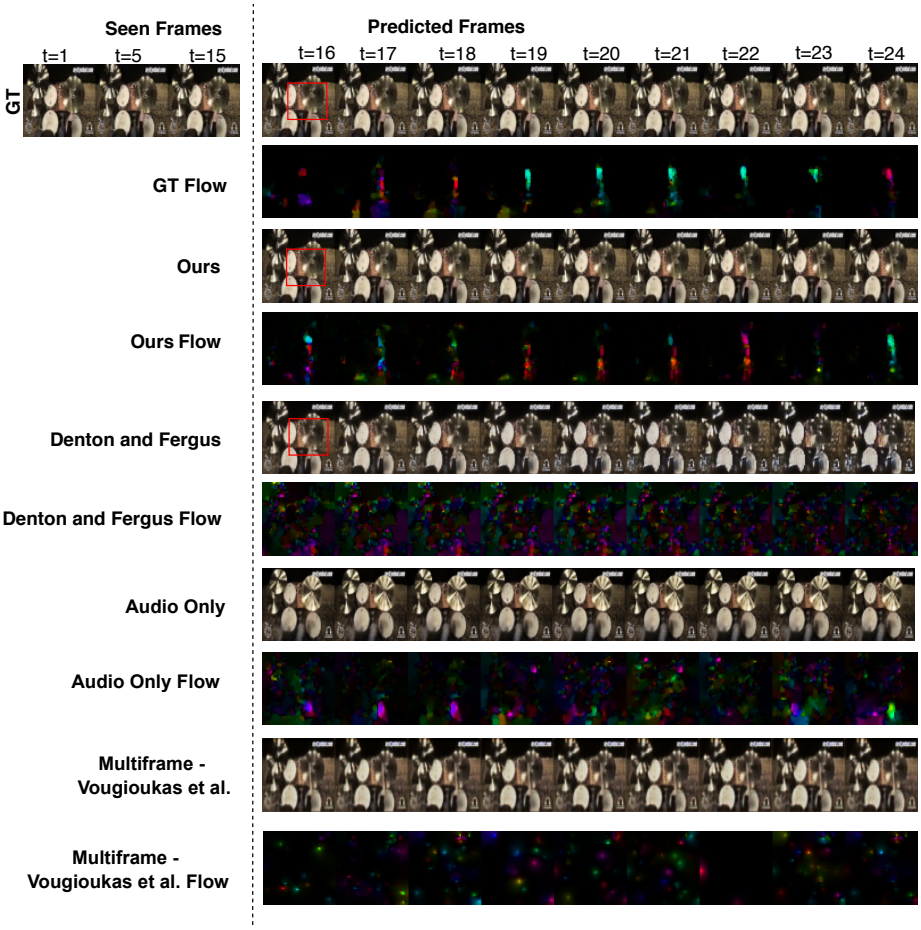


Fig. 47: Sample generations on the AudioSet-Drums dataset by our method vis-á-vis other baselines and optical flows across frames. The red square denotes regions of high motion.

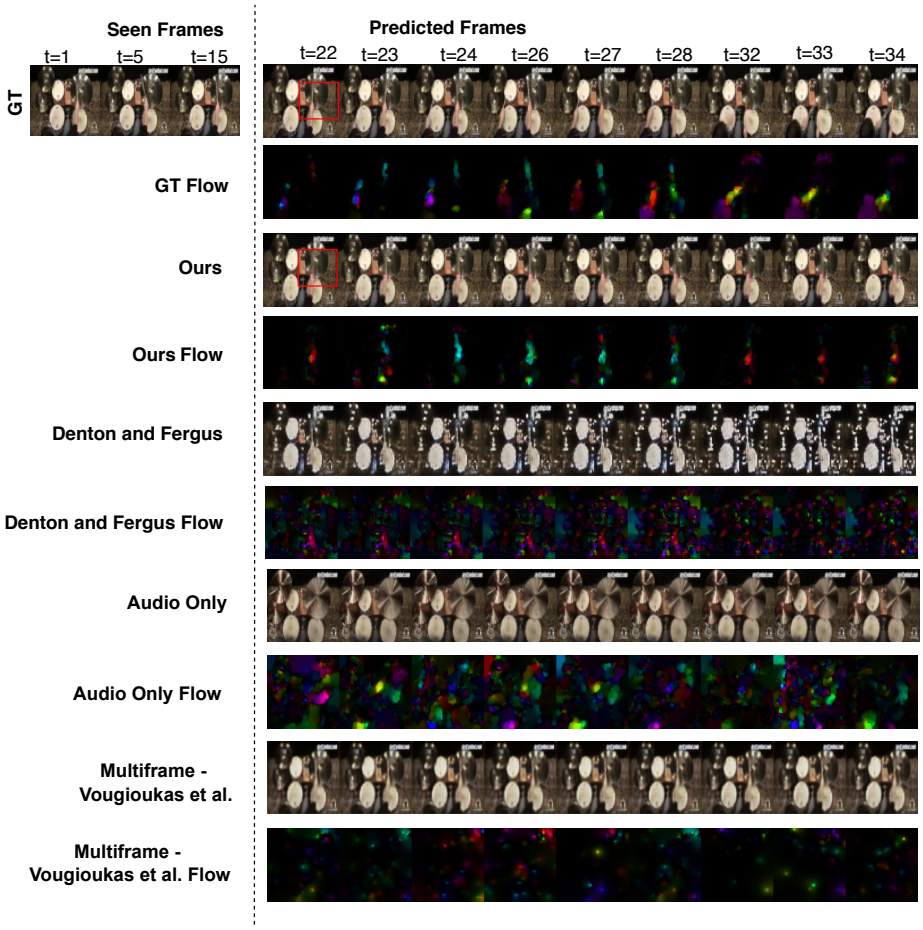


Fig. 48: Sample generations on the AudioSet-Drums dataset by our method vis-á-vis other baselines and optical flows across frames. The red square denotes regions of high motion.

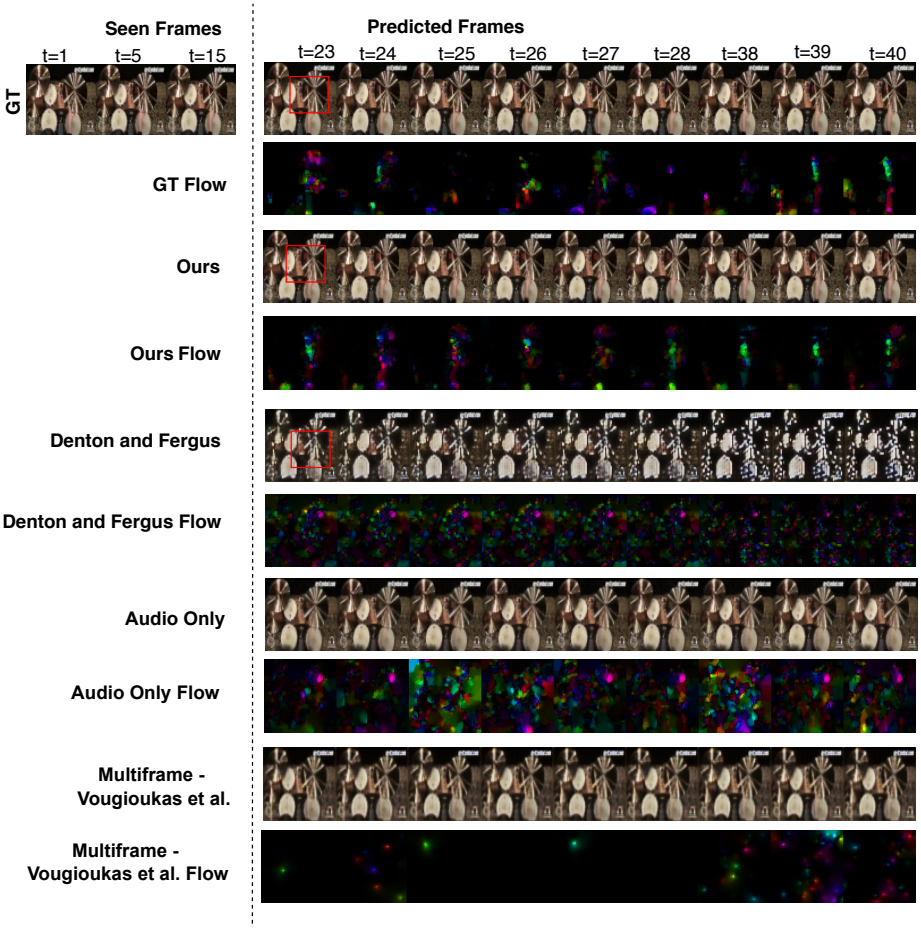


Fig. 49: Sample generations on the AudioSet-Drums dataset by our method vis-á-vis other baselines and optical flows across frames. The red square denotes regions of high motion.

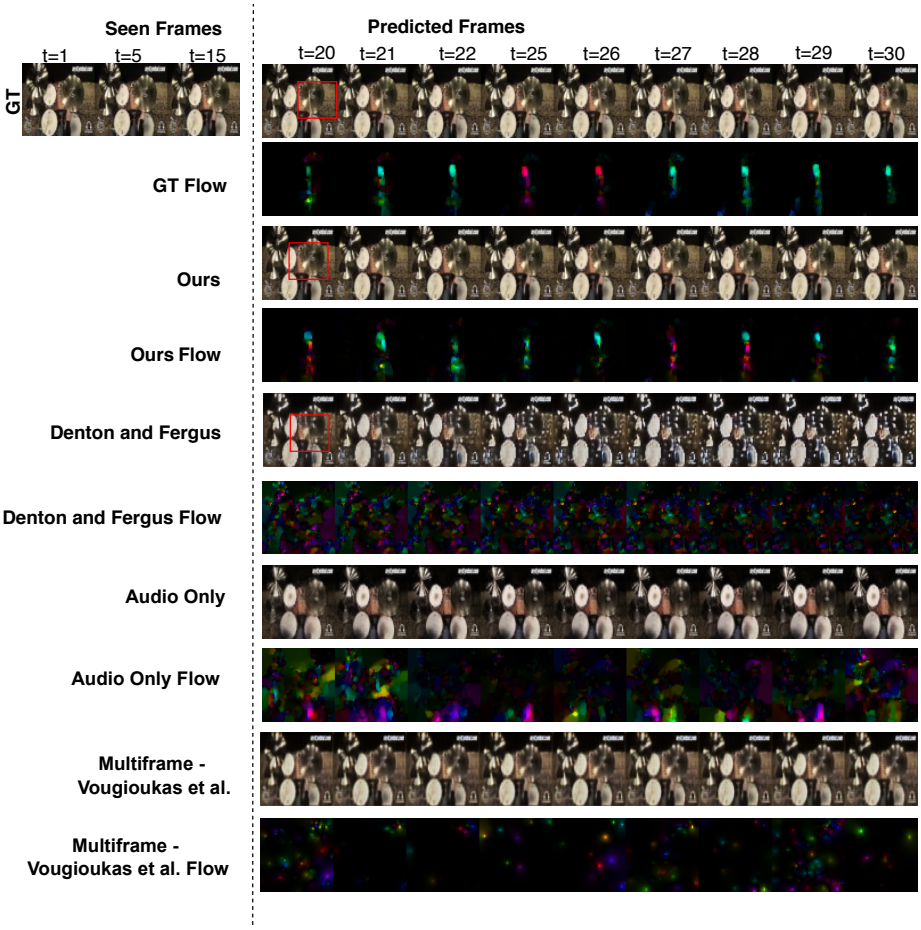


Fig. 50: Sample generations on the AudioSet-Drums dataset by our method vis-á-vis other baselines and optical flows across frames. The red square denotes regions of high motion.

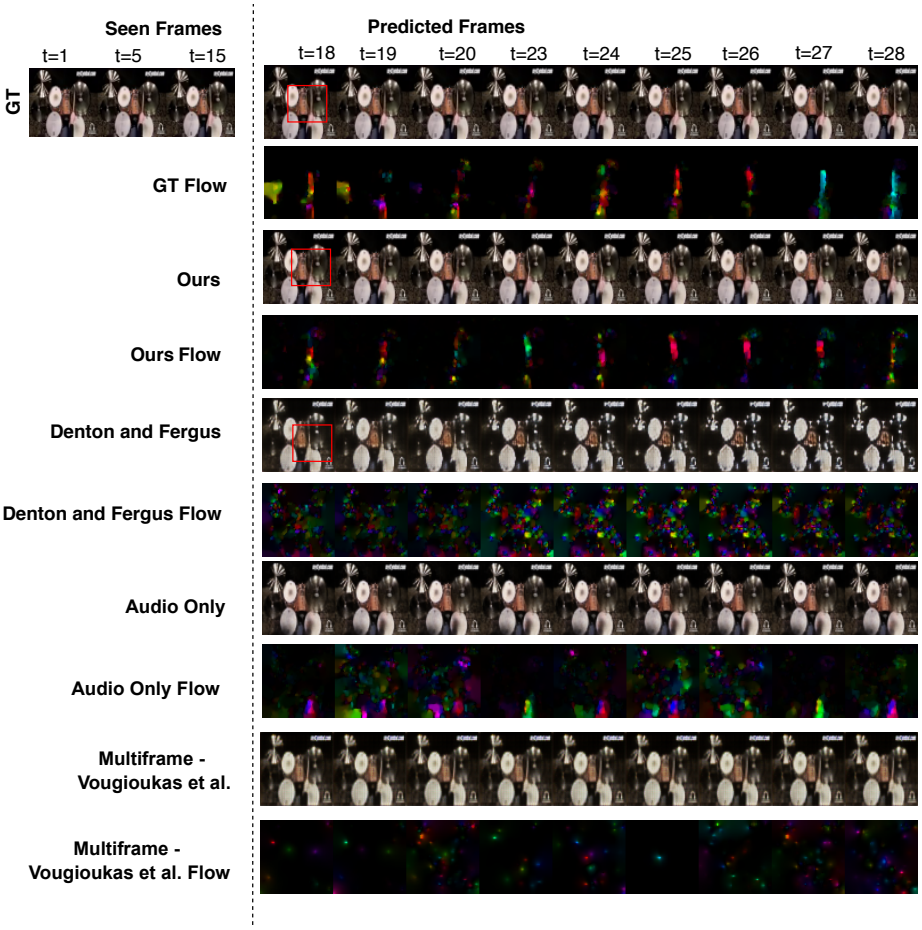


Fig. 51: Sample generations on the AudioSet-Drums dataset by our method vis-á-vis other baselines and optical flows across frames. The red square denotes regions of high motion.

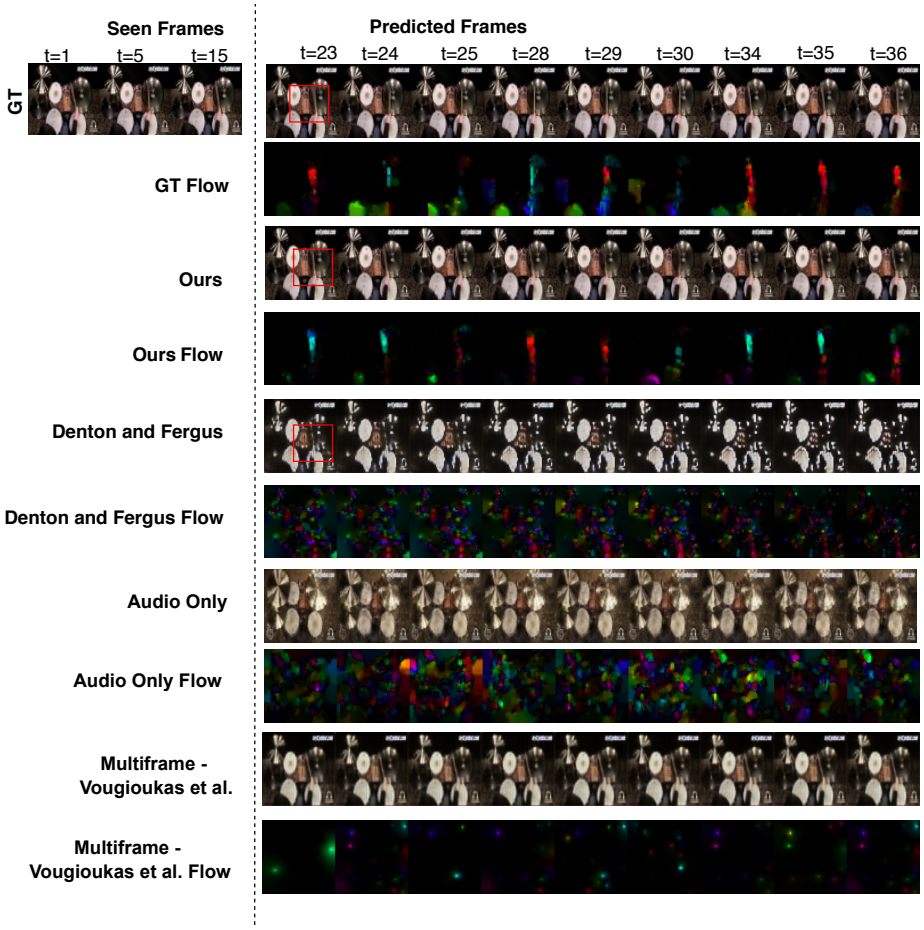


Fig. 52: Sample generations on the AudioSet-Drums dataset by our method vis-á-vis other baselines and optical flows across frames. The red square denotes regions of high motion.

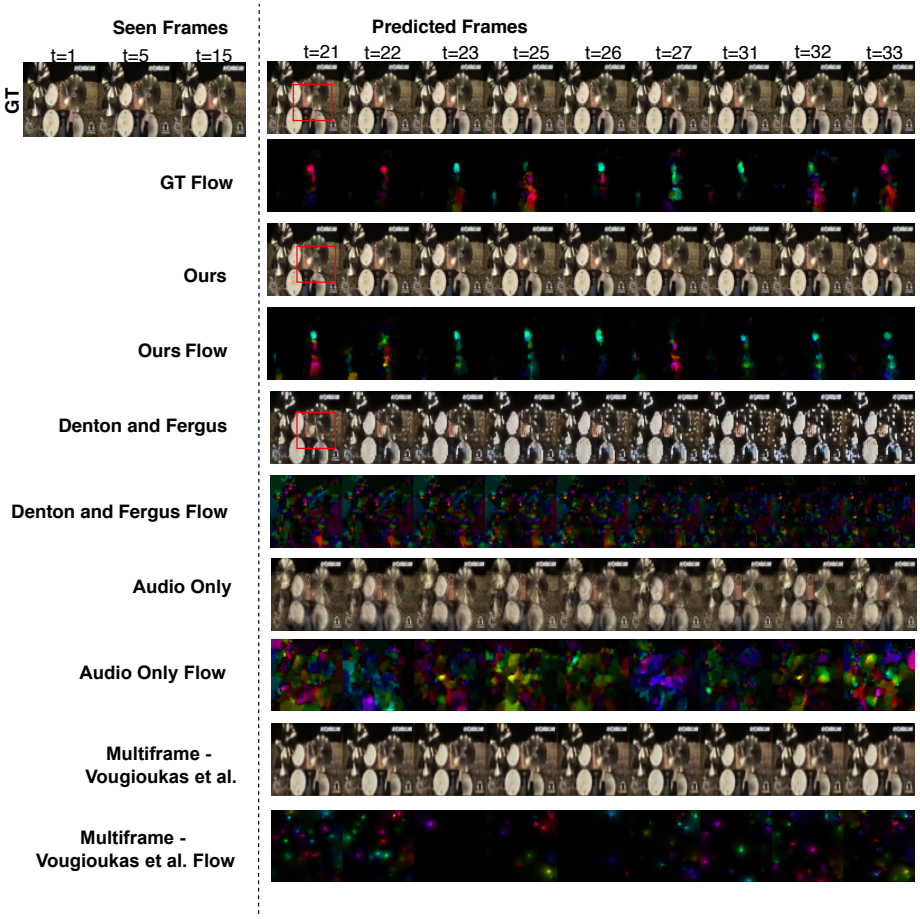


Fig. 53: Sample generations on the AudioSet-Drums dataset by our method vis-á-vis other baselines and optical flows across frames. The red square denotes regions of high motion.

7.1 Diverse Sample Generations

In Figures 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68, 69, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90, 91, 92, 93, 94, 95, 96, 97, 98 we present qualitative visualizations of a set of diverse generations for every sample across all datasets. The green box highlights frames which are noticeably distinct across samples, underscoring the variety of the generated samples for both synthetic and real-world datasets.

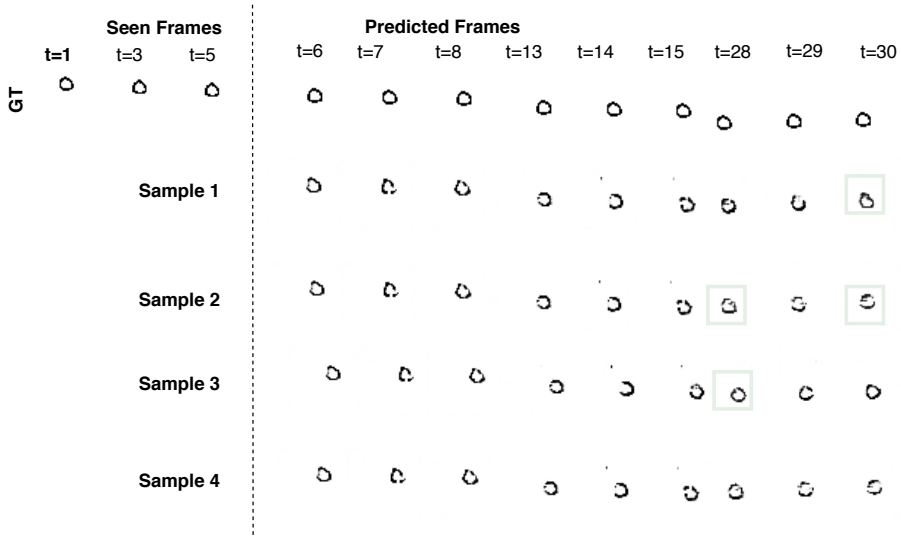


Fig. 54: Diverse sample generations on the M3SO-NB dataset by our method. The green square highlights frames where noticeable differences are observed across samples.

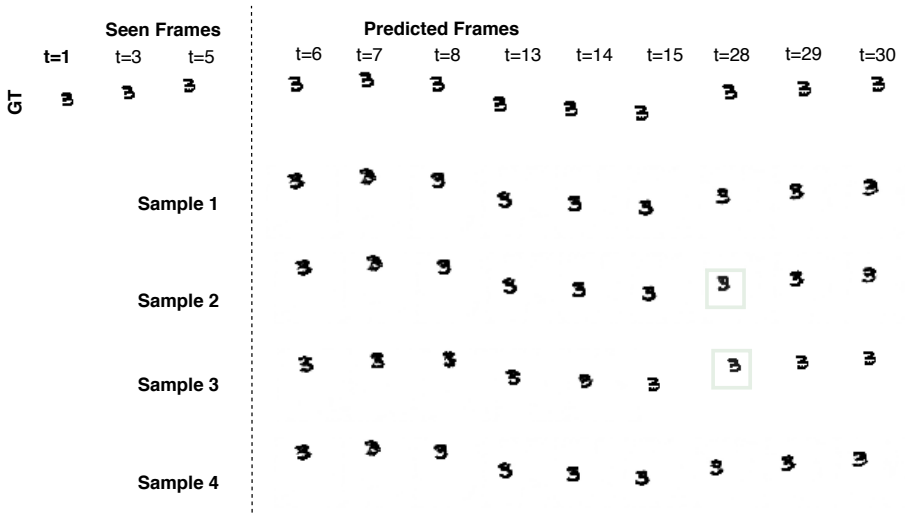


Fig. 55: Diverse sample generations on the M3SO-NB dataset by our method. The green square highlights frames where noticeable differences are observed across samples.

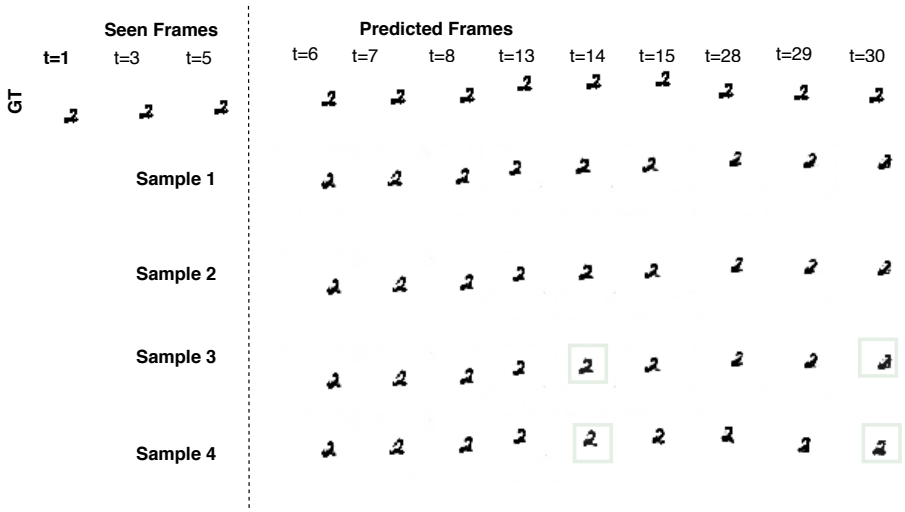


Fig. 56: Diverse sample generations on the M3SO-NB dataset by our method. The green square highlights frames where noticeable differences are observed across samples.

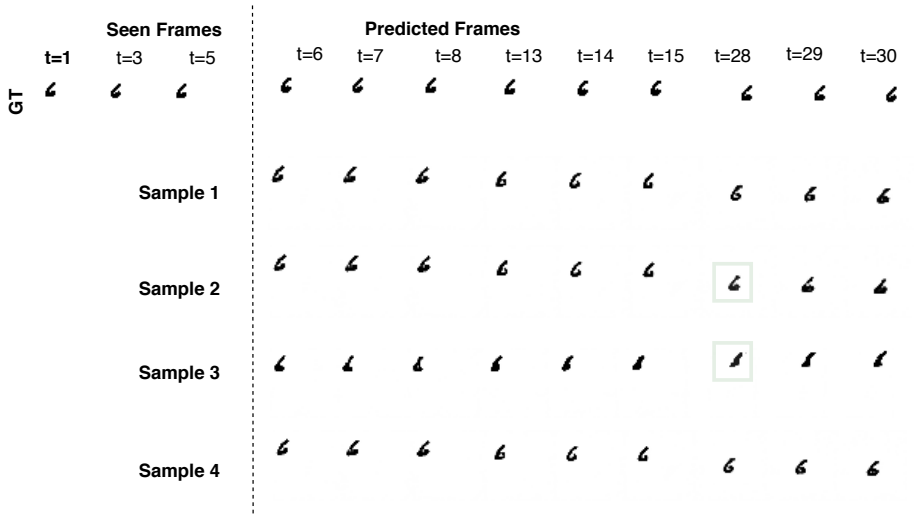


Fig. 57: Diverse sample generations on the M3SO-NB dataset by our method. The green square highlights frames where noticeable differences are observed across samples.

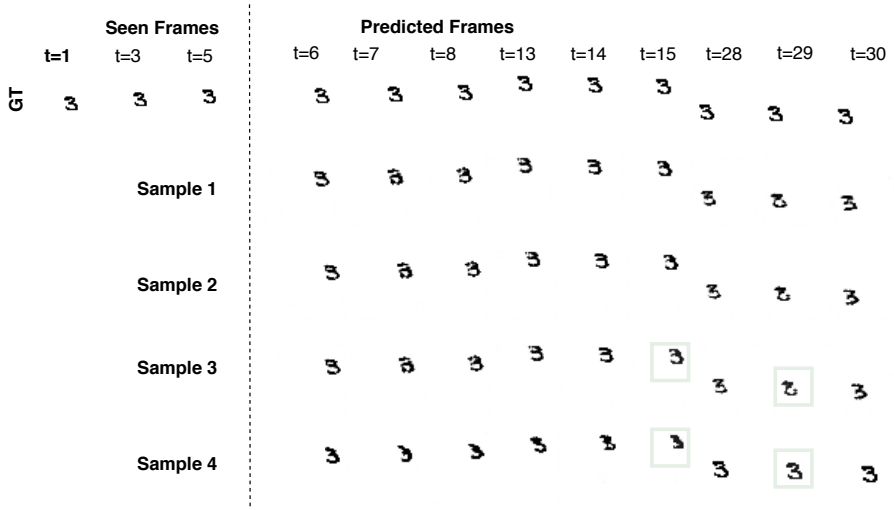


Fig. 58: Diverse sample generations on the M3SO-NB dataset by our method. The green square highlights frames where noticeable differences are observed across samples.

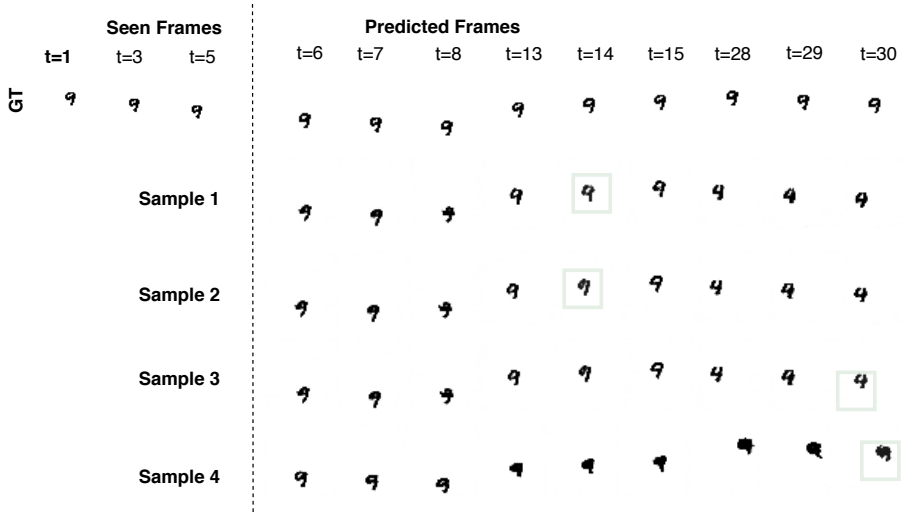


Fig. 59: Diverse sample generations on the M3SO-NB dataset by our method. The green square highlights frames where noticeable differences are observed across samples.

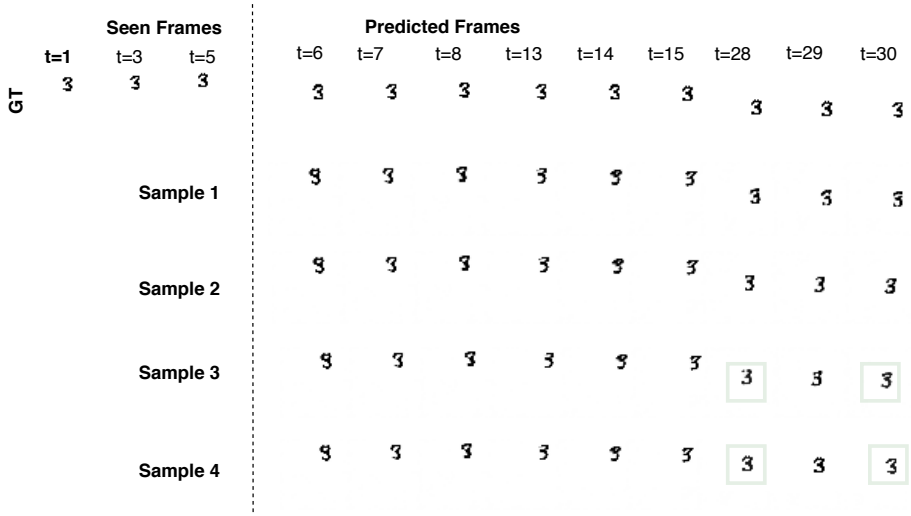


Fig. 60: Diverse sample generations on the M3SO-NB dataset by our method. The green square highlights frames where noticeable differences are observed across samples.

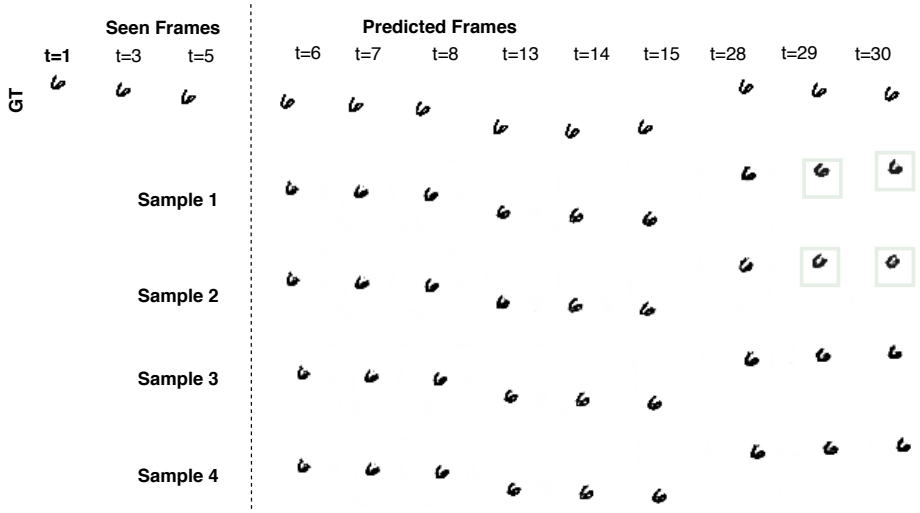


Fig. 61: Diverse sample generations on the M3SO-NB dataset by our method. The green square highlights frames where noticeable differences are observed across samples.

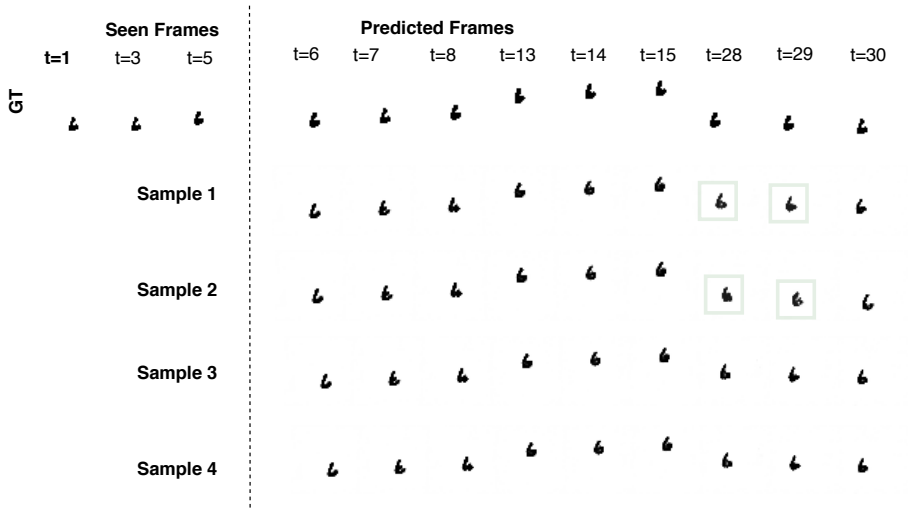


Fig. 62: Diverse sample generations on the M3SO-NB dataset by our method. The green square highlights frames where noticeable differences are observed across samples.

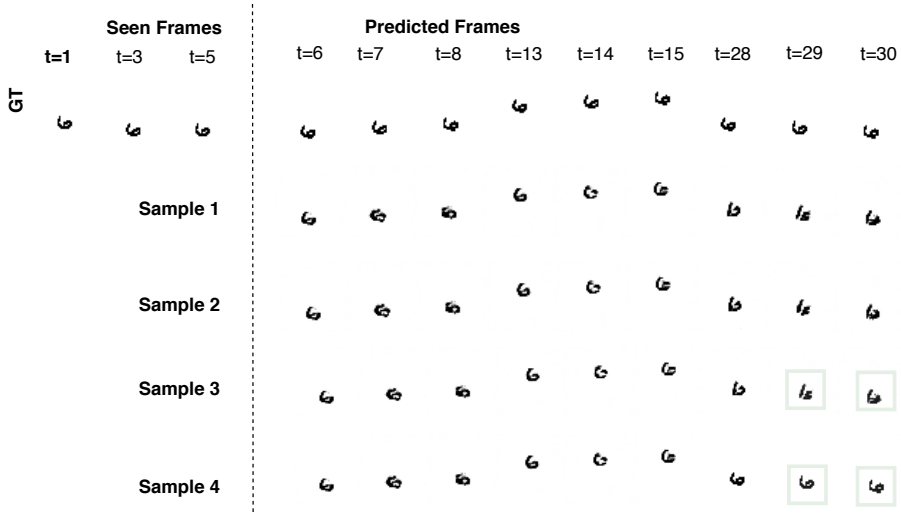


Fig. 63: Diverse sample generations on the M3SO-NB dataset by our method. The green square highlights frames where noticeable differences are observed across samples.

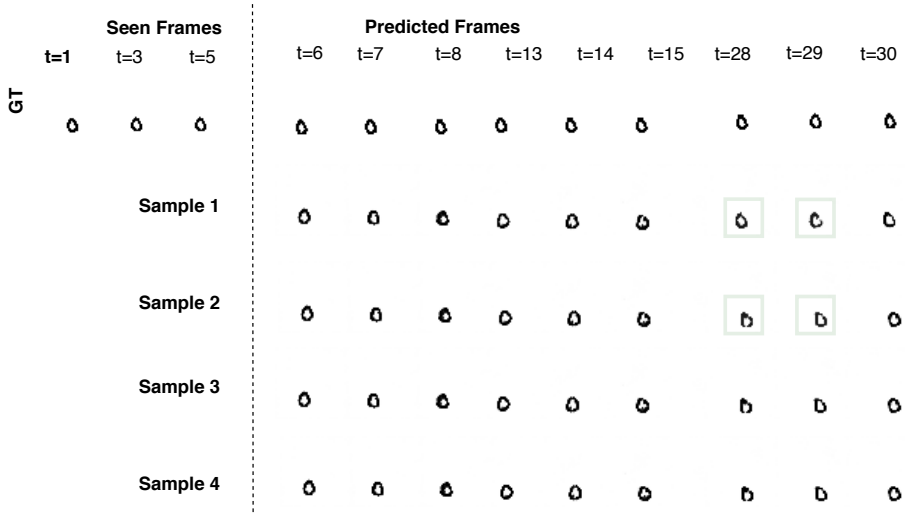


Fig. 64: Diverse sample generations on the M3SO-NB dataset by our method. The green square highlights frames where noticeable differences are observed across samples.

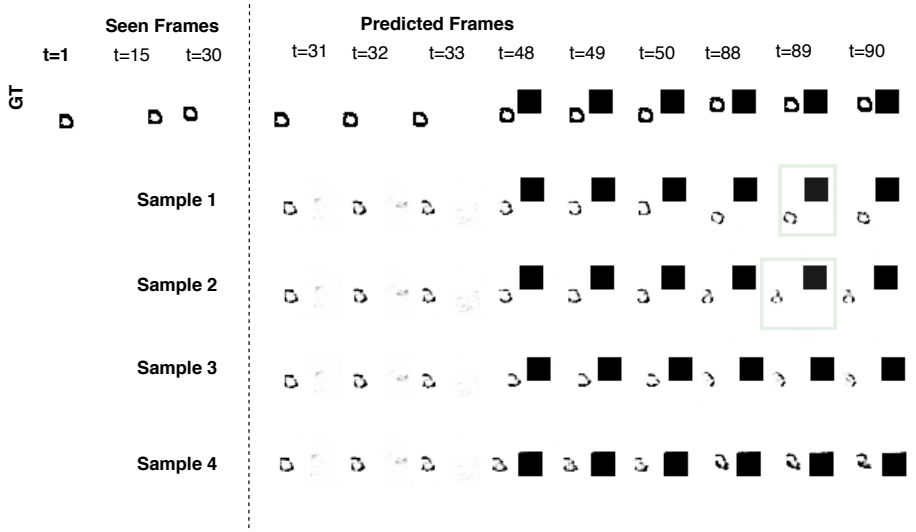


Fig. 65: Diverse sample generations on the Multimodal MovingMNIST with Surprise Obstacle dataset by our method. The green square highlights frames where noticeable differences are observed across samples.

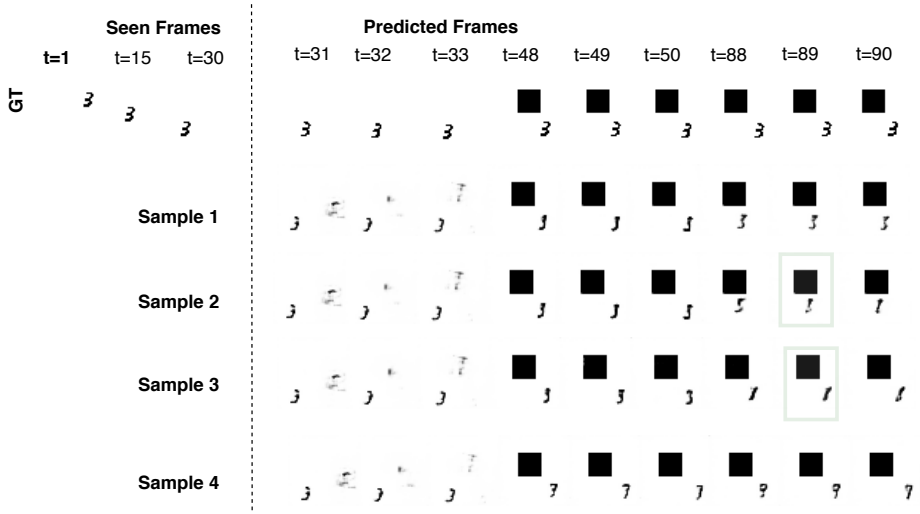


Fig. 66: Diverse sample generations on the Multimodal MovingMNIST with Surprise Obstacle dataset by our method. The green square highlights frames where noticeable differences are observed across samples.

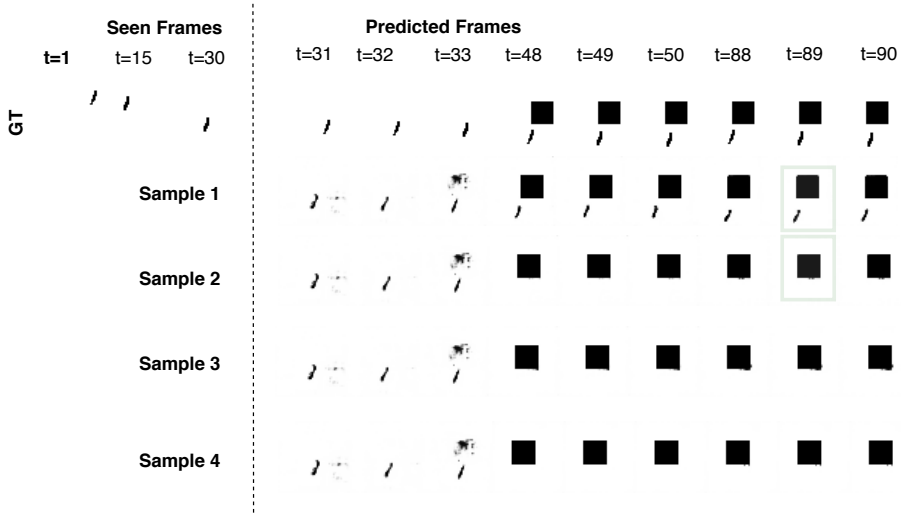


Fig. 67: Diverse sample generations on the Multimodal MovingMNIST with Surprise Obstacle dataset by our method. The green square highlights frames where noticeable differences are observed across samples.

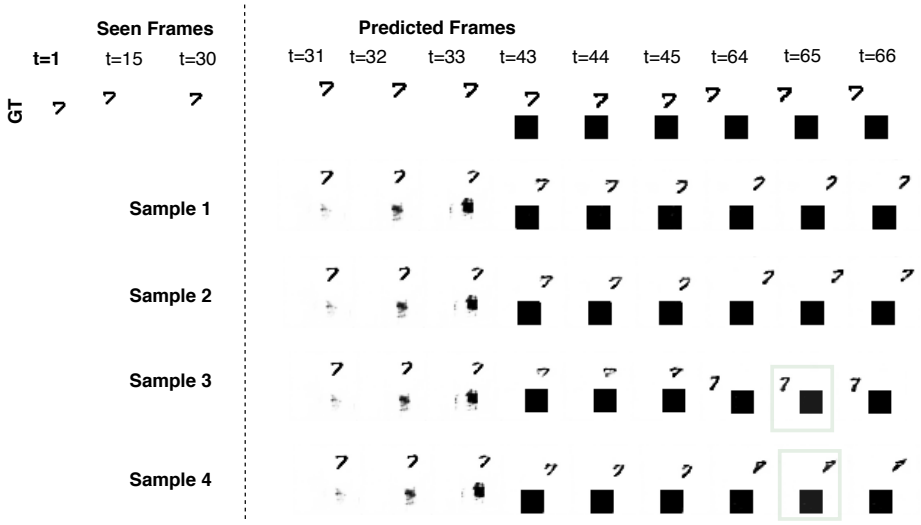


Fig. 68: Diverse sample generations on the Multimodal MovingMNIST with Surprise Obstacle dataset by our method. The green square highlights frames where noticeable differences are observed across samples.

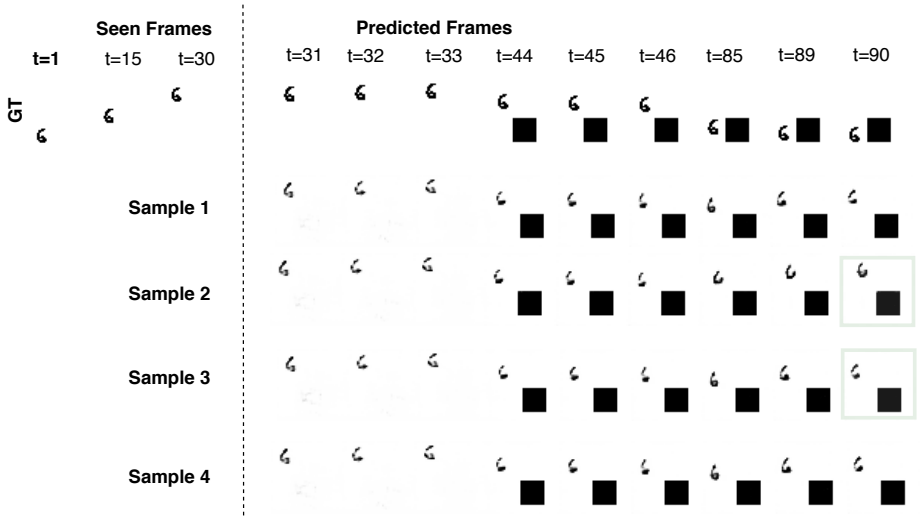


Fig. 69: Diverse sample generations on the Multimodal MovingMNIST with Surprise Obstacle dataset by our method. The green square highlights frames where noticeable differences are observed across samples.

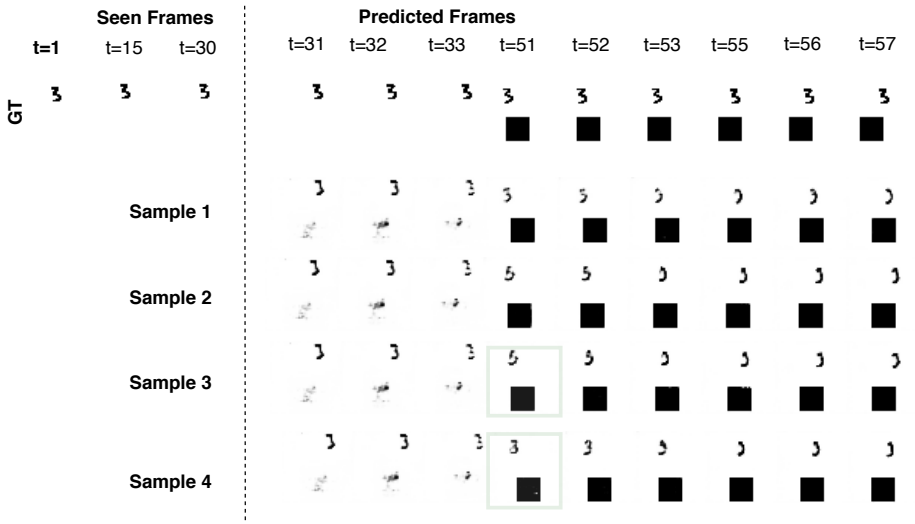


Fig. 70: Diverse sample generations on the Multimodal MovingMNIST with Surprise Obstacle dataset by our method. The green square highlights frames where noticeable differences are observed across samples.

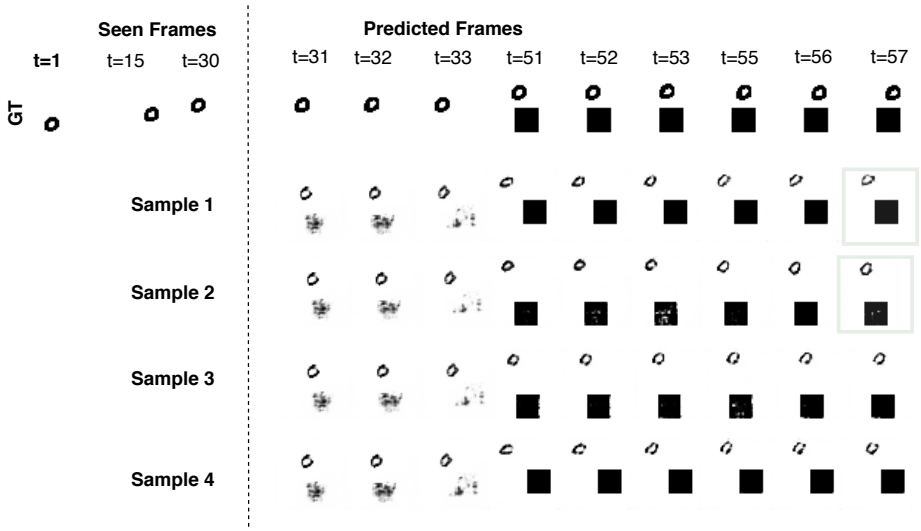


Fig. 71: Diverse sample generations on the Multimodal MovingMNIST with Surprise Obstacle dataset by our method. The green square highlights frames where noticeable differences are observed across samples.

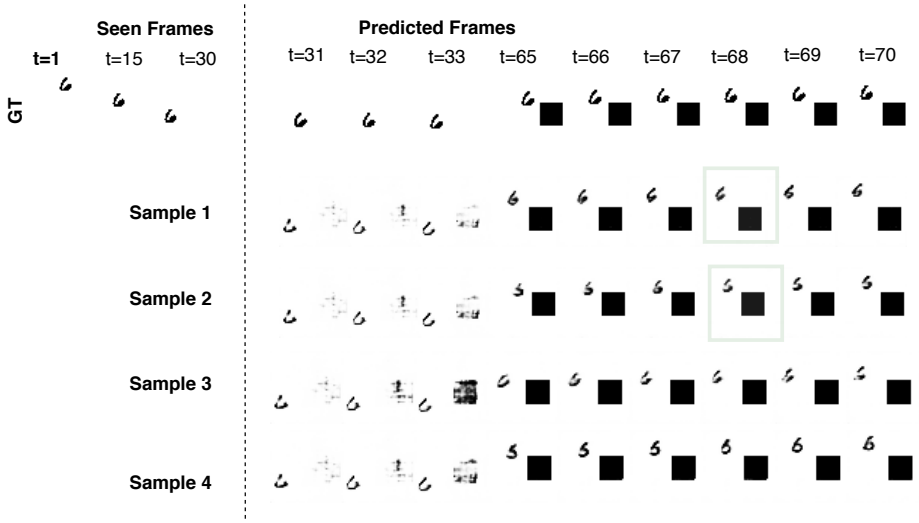


Fig. 72: Diverse sample generations on the Multimodal MovingMNIST with Surprise Obstacle dataset by our method. The green square highlights frames where noticeable differences are observed across samples.

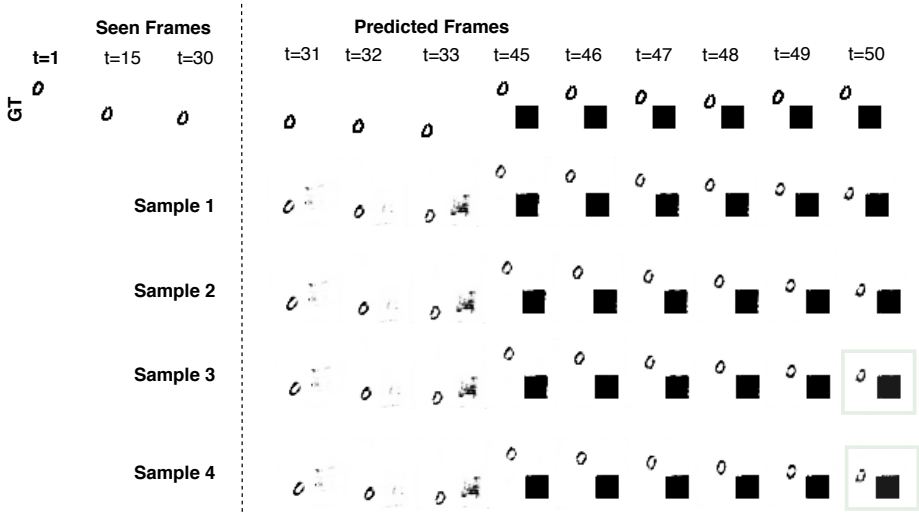


Fig. 73: Diverse sample generations on the Multimodal MovingMNIST with Surprise Obstacle dataset by our method. The green square highlights frames where noticeable differences are observed across samples.

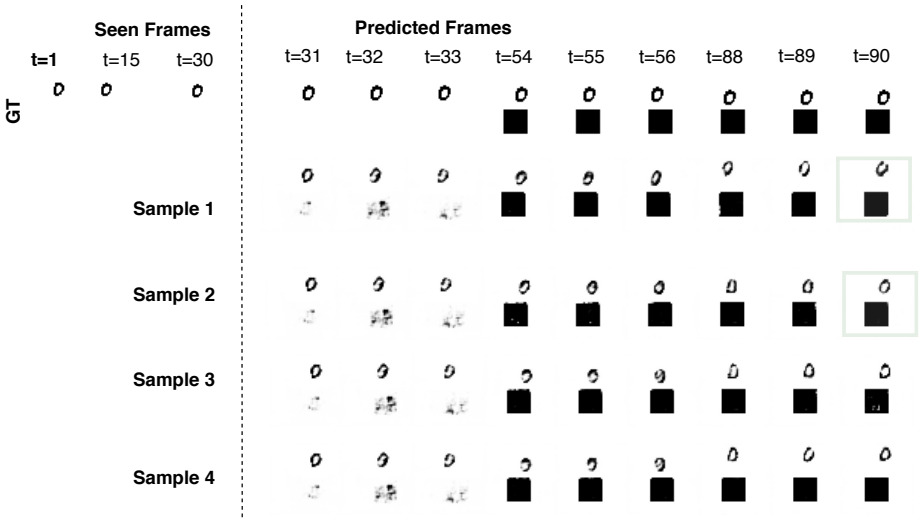


Fig. 74: Diverse sample generations on the Multimodal MovingMNIST with Surprise Obstacle dataset by our method. The green square highlights frames where noticeable differences are observed across samples.

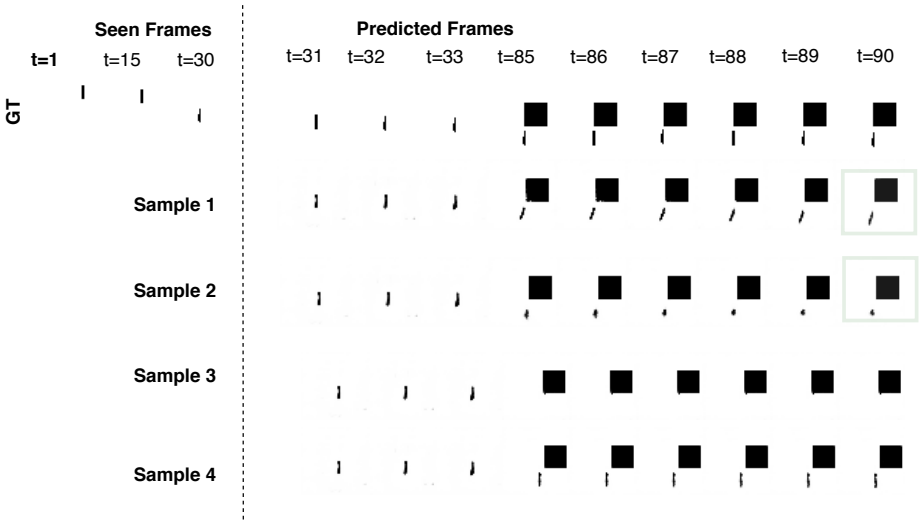


Fig. 75: Diverse sample generations on the Multimodal MovingMNIST with Surprise Obstacle dataset by our method. The green square highlights frames where noticeable differences are observed across samples.

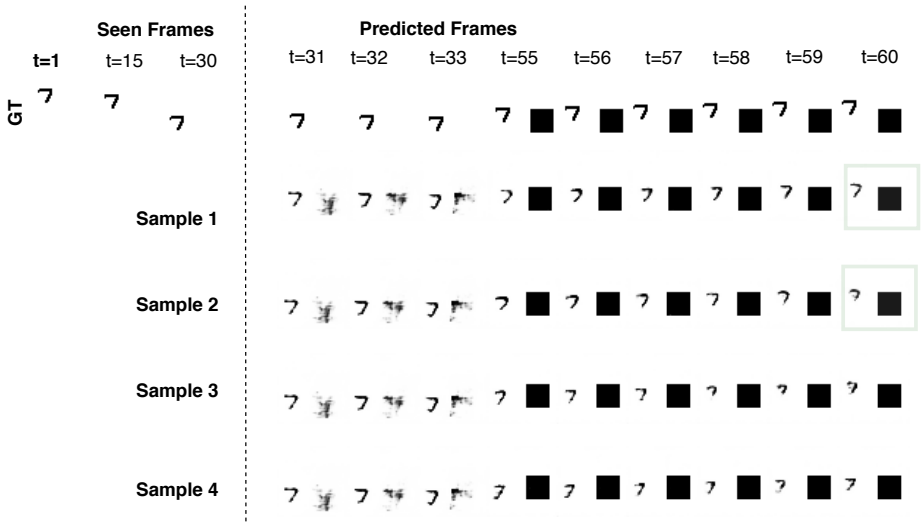


Fig. 76: Diverse sample generations on the Multimodal MovingMNIST with Surprise Obstacle dataset by our method. The green square highlights frames where noticeable differences are observed across samples.

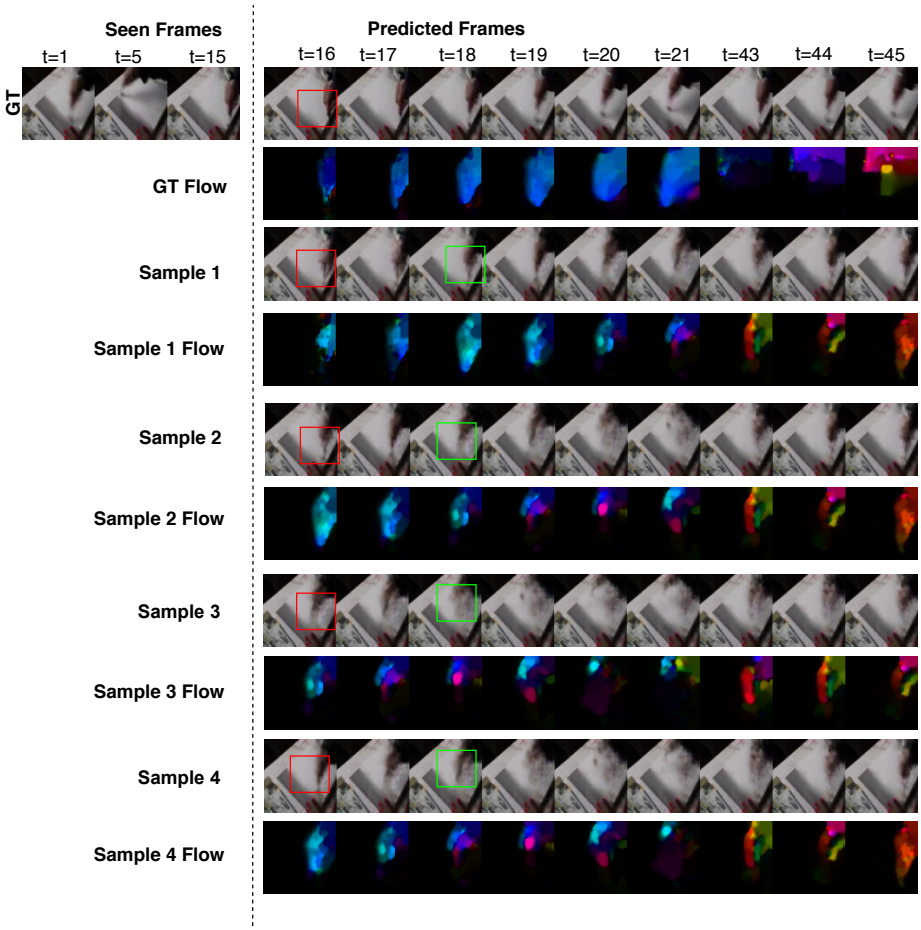


Fig. 77: Diverse sample generations on the YouTube Painting dataset by our method along with the optical flows between frames. The red square denotes regions of high motion, while the green square highlights frames where noticeable differences are observed across samples.

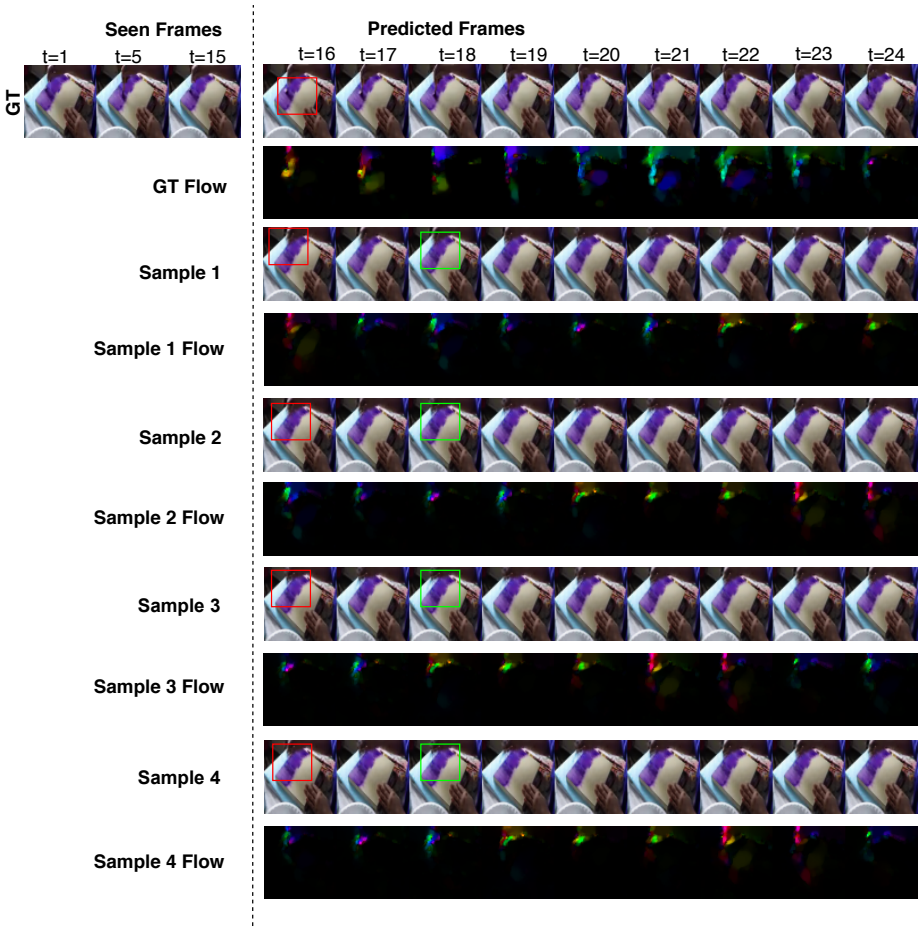


Fig. 78: Diverse sample generations on the YouTube Painting dataset by our method along with the optical flows between frames. The red square denotes regions of high motion, while the green square highlights frames where noticeable differences are observed across samples.

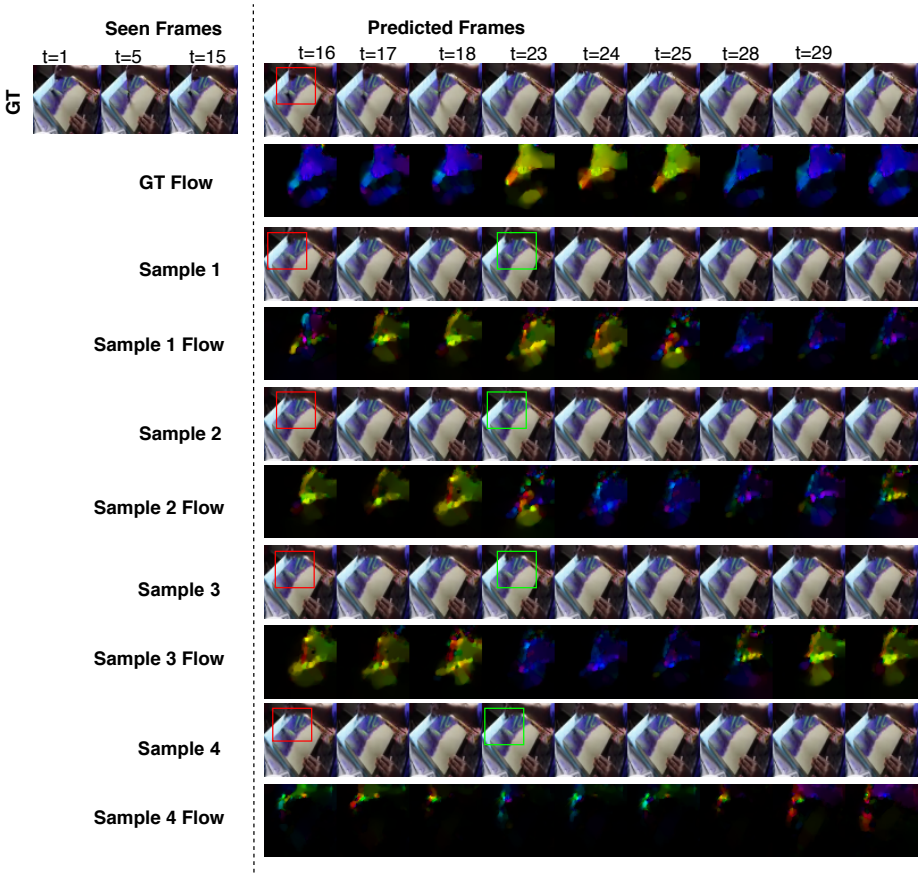


Fig. 79: Diverse sample generations on the YouTube Painting dataset by our method along with the optical flows between frames. The red square denotes regions of high motion, while the green square highlights frames where noticeable differences are observed across samples.

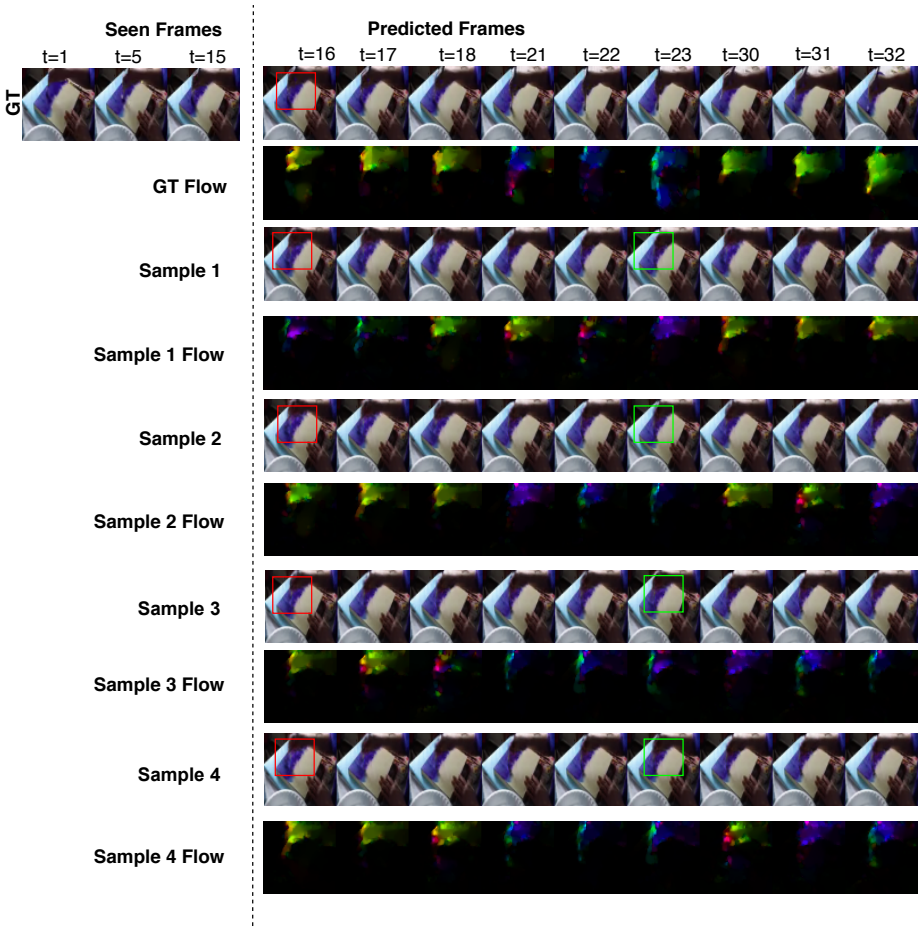


Fig. 80: Diverse sample generations on the YouTube Painting dataset by our method along with the optical flows between frames. The red square denotes regions of high motion, while the green square highlights frames where noticeable differences are observed across samples.

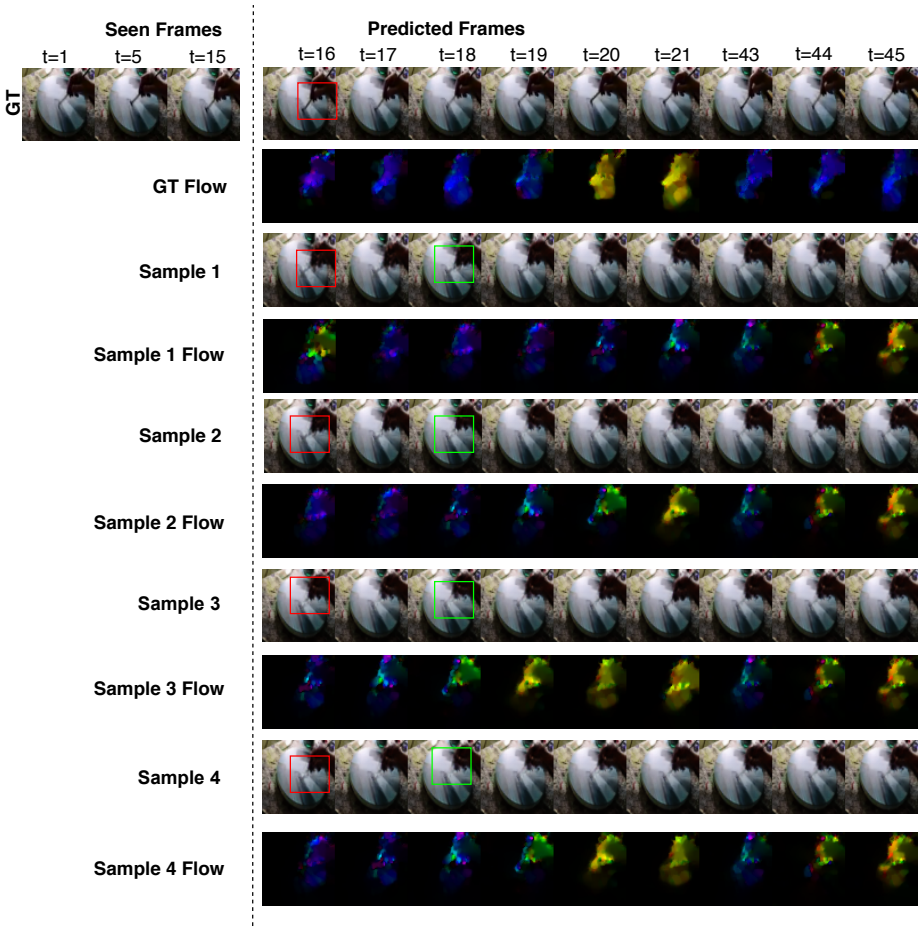


Fig. 81: Diverse sample generations on the YouTube Painting dataset by our method along with the optical flows between frames. The red square denotes regions of high motion, while the green square highlights frames where noticeable differences are observed across samples.

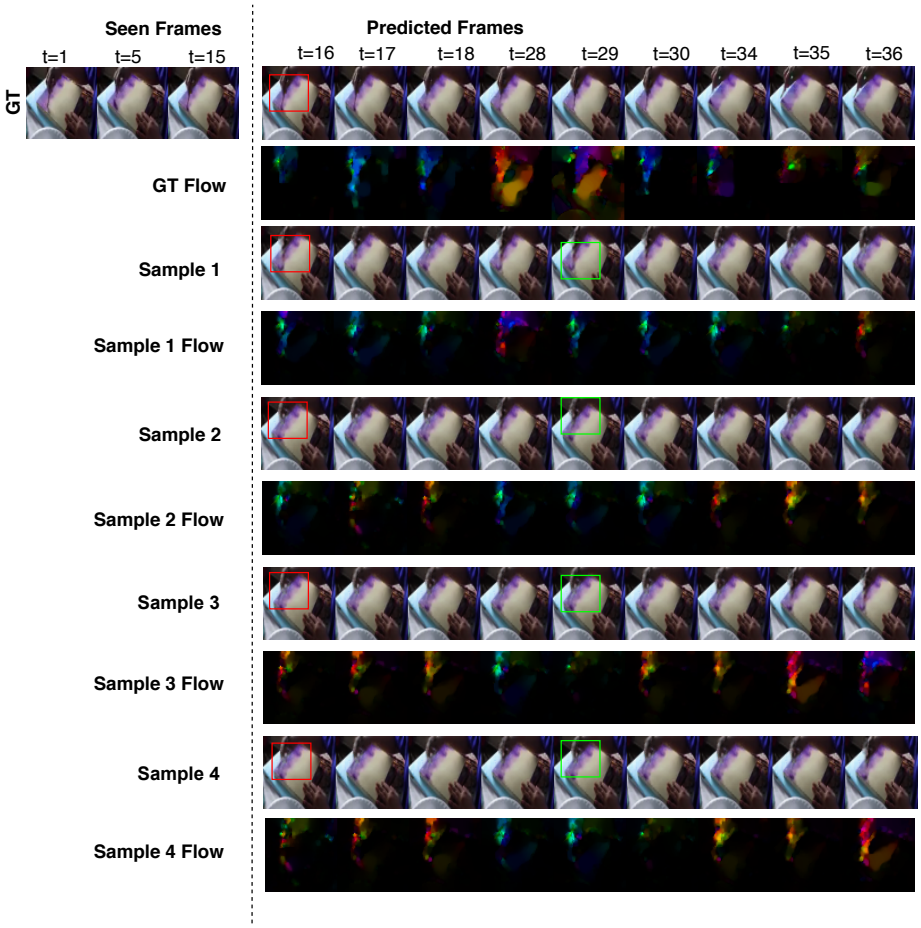


Fig. 82: Diverse sample generations on the YouTube Painting dataset by our method along with the optical flows between frames. The red square denotes regions of high motion, while the green square highlights frames where noticeable differences are observed across samples.

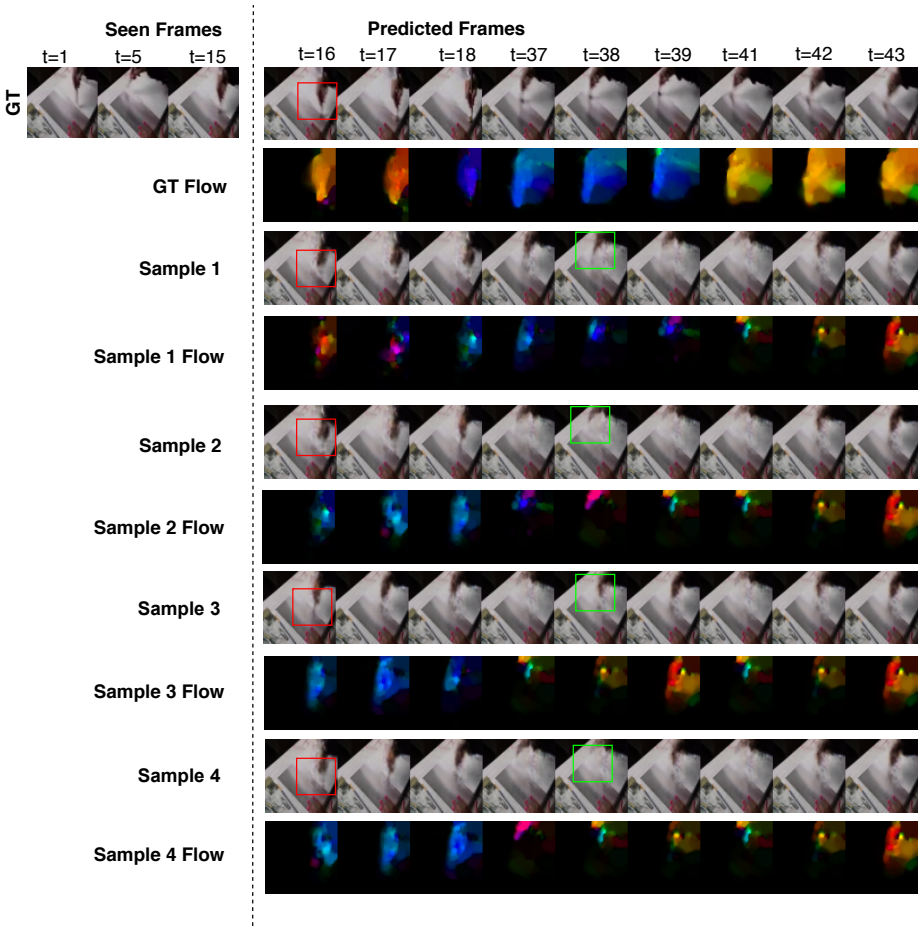


Fig. 83: Diverse sample generations on the YouTube Painting dataset by our method along with the optical flows between frames. The red square denotes regions of high motion, while the green square highlights frames where noticeable differences are observed across samples.

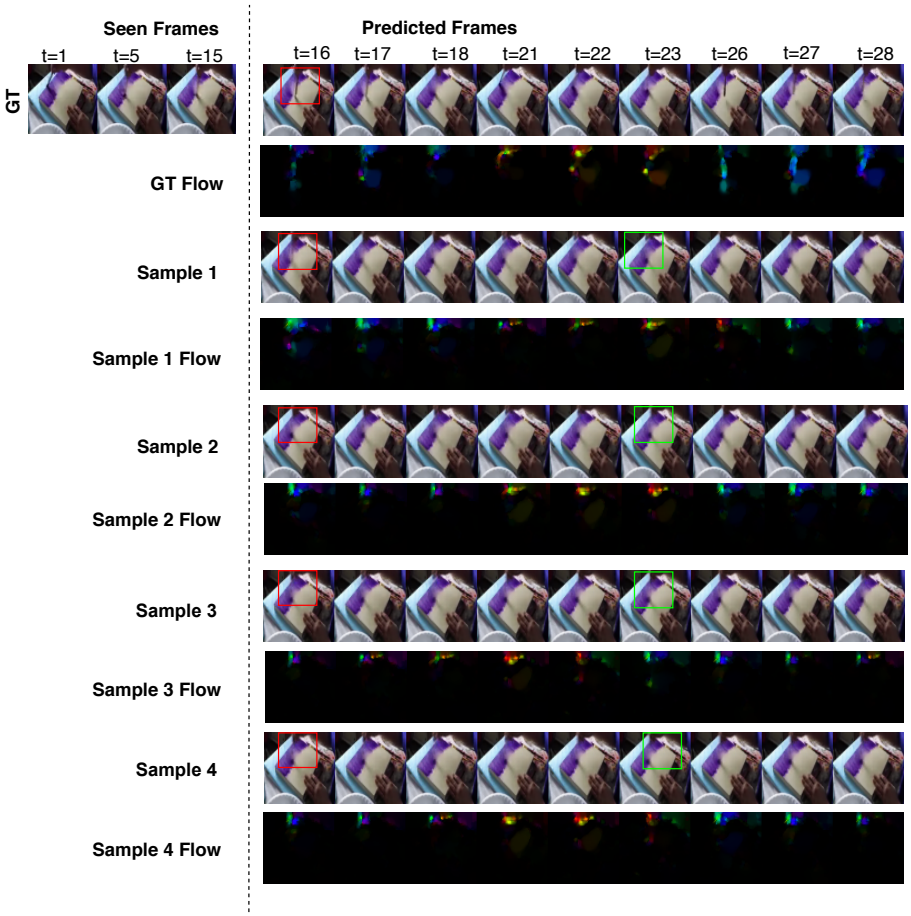


Fig. 84: Diverse sample generations on the YouTube Painting dataset by our method along with the optical flows between frames. The red square denotes regions of high motion, while the green square highlights frames where noticeable differences are observed across samples.

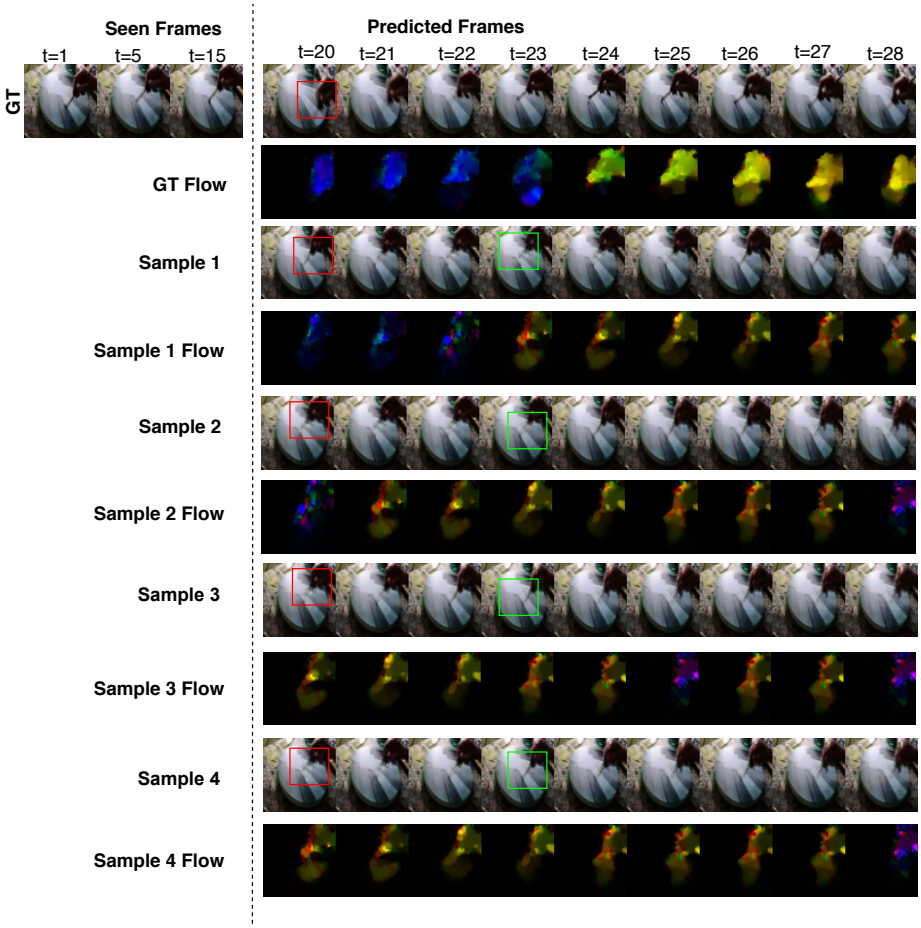


Fig. 85: Diverse sample generations on the YouTube Painting dataset by our method along with the optical flows between frames. The red square denotes regions of high motion, while the green square highlights frames where noticeable differences are observed across samples.

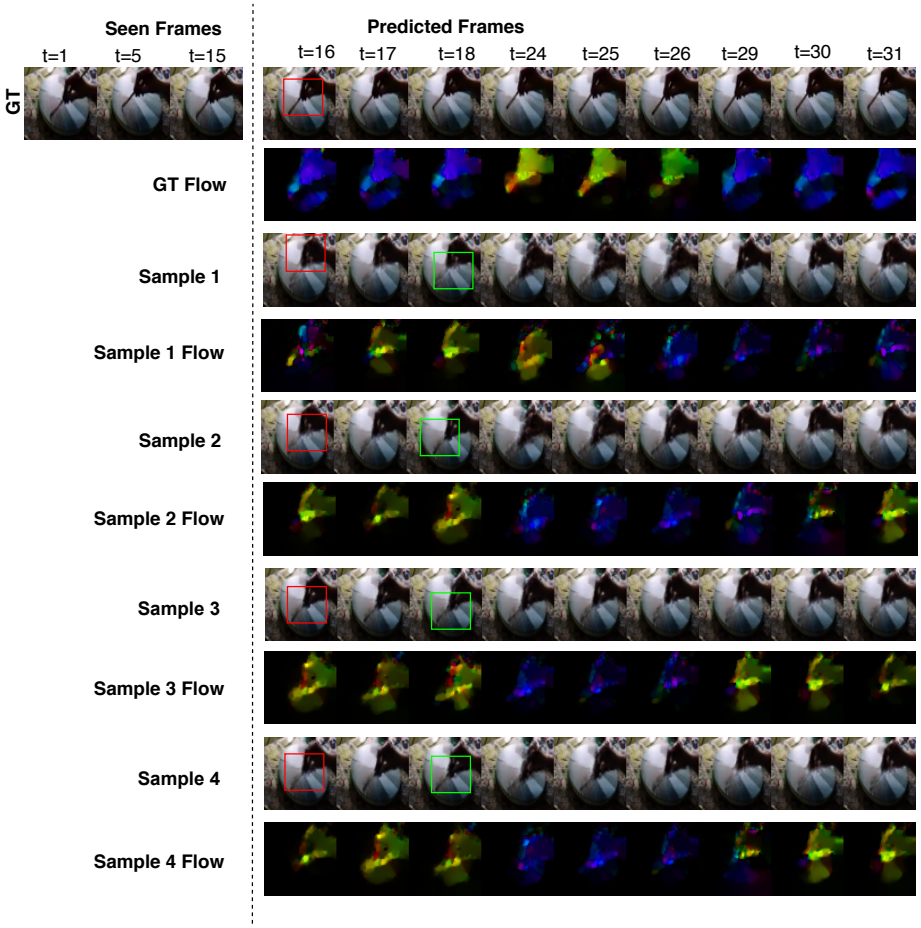


Fig. 86: Diverse sample generations on the YouTube Painting dataset by our method along with the optical flows between frames. The red square denotes regions of high motion, while the green square highlights frames where noticeable differences are observed across samples.

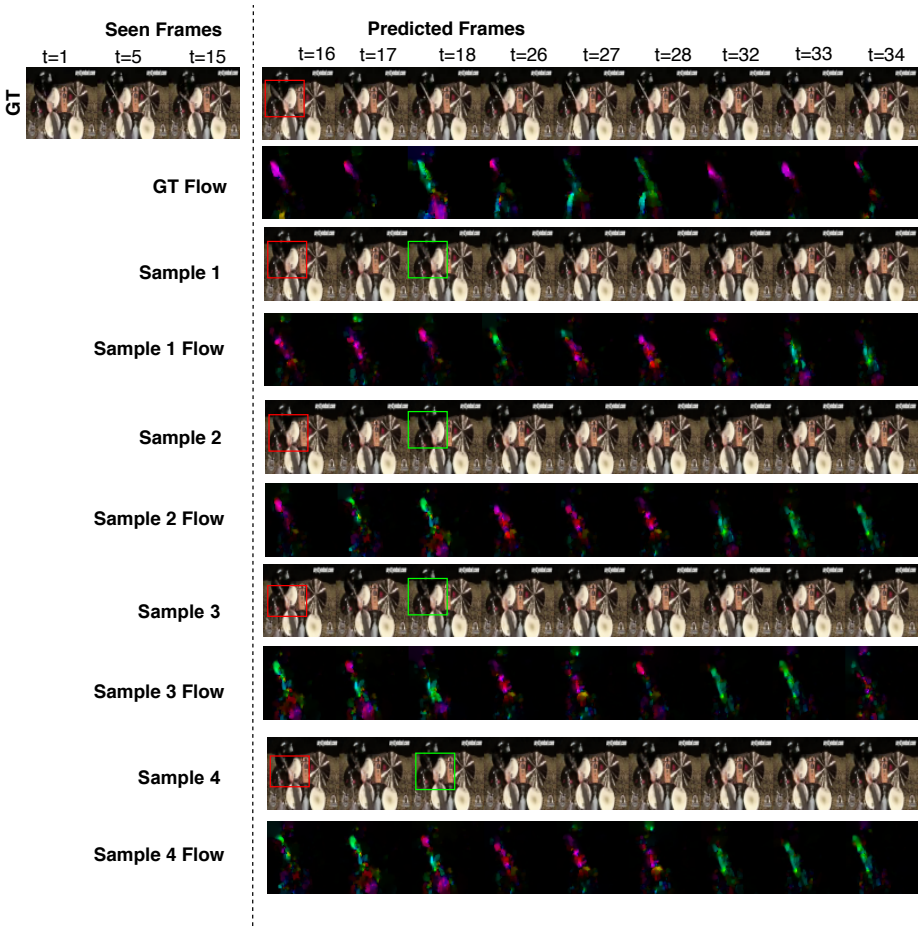


Fig. 87: Diverse sample generations on the AudioSet Drums dataset by our method along with the optical flows between frames. The red square denotes regions of high motion, while the green square highlights frames where noticeable differences are observed across samples.

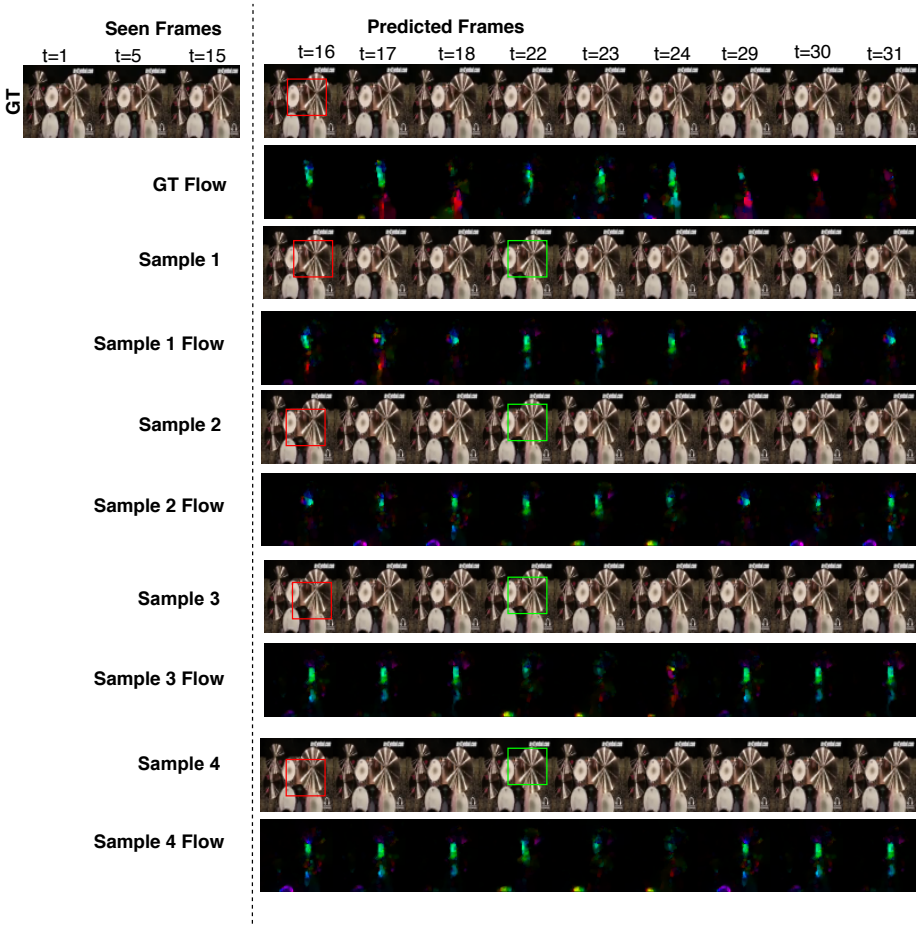


Fig. 88: Diverse sample generations on the AudioSet Drums dataset by our method along with the optical flows between frames. The red square denotes regions of high motion, while the green square highlights frames where noticeable differences are observed across samples.

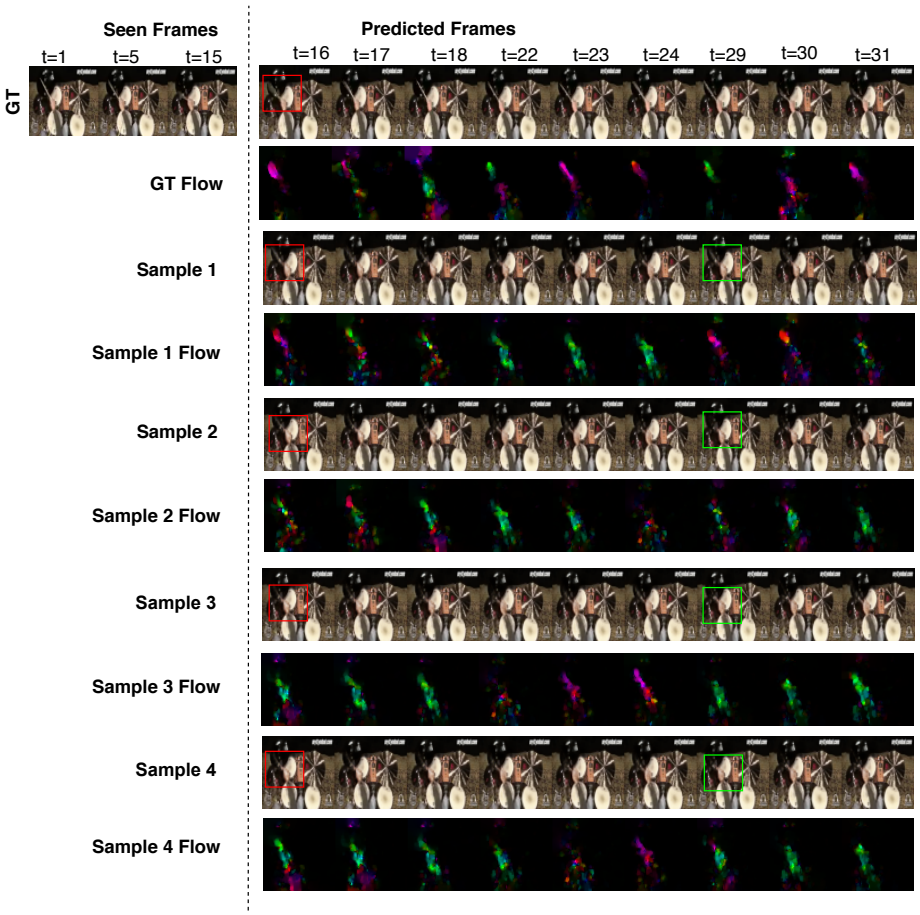


Fig. 89: Diverse sample generations on the AudioSet Drums dataset by our method along with the optical flows between frames. The red square denotes regions of high motion, while the green square highlights frames where noticeable differences are observed across samples.

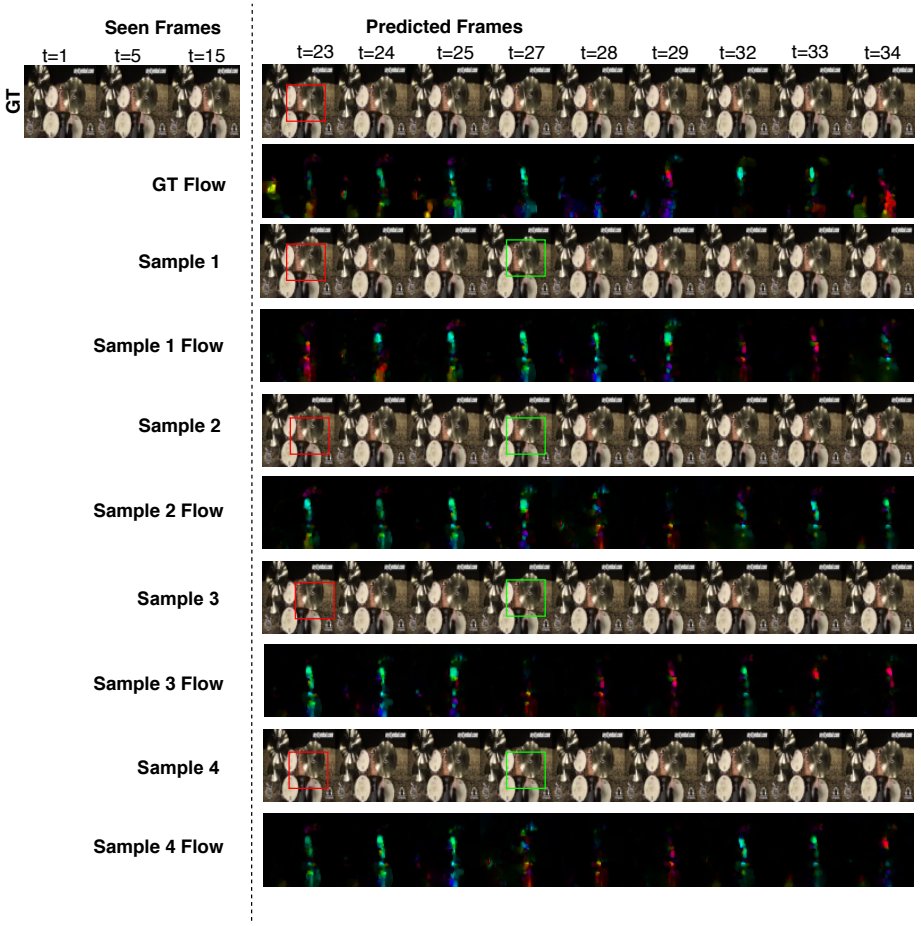


Fig. 90: Diverse sample generations on the AudioSet Drums dataset by our method along with the optical flows between frames. The red square denotes regions of high motion, while the green square highlights frames where noticeable differences are observed across samples.

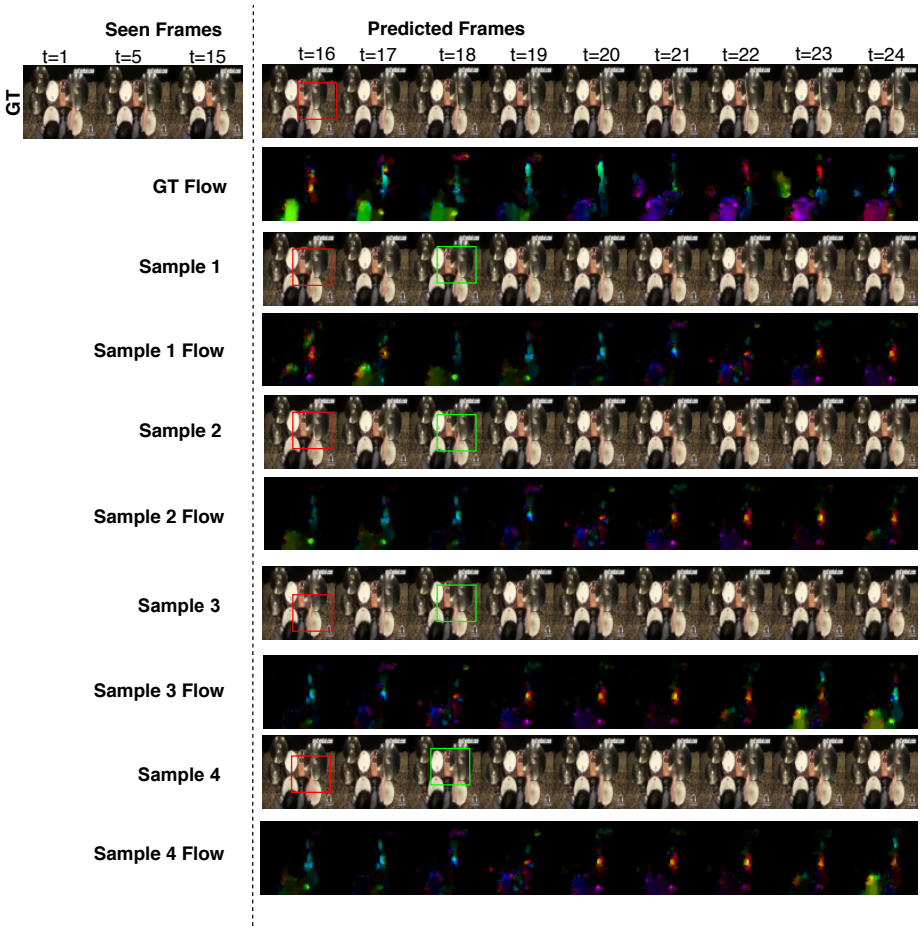


Fig. 91: Diverse sample generations on the AudioSet Drums dataset by our method along with the optical flows between frames. The red square denotes regions of high motion, while the green square highlights frames where noticeable differences are observed across samples.

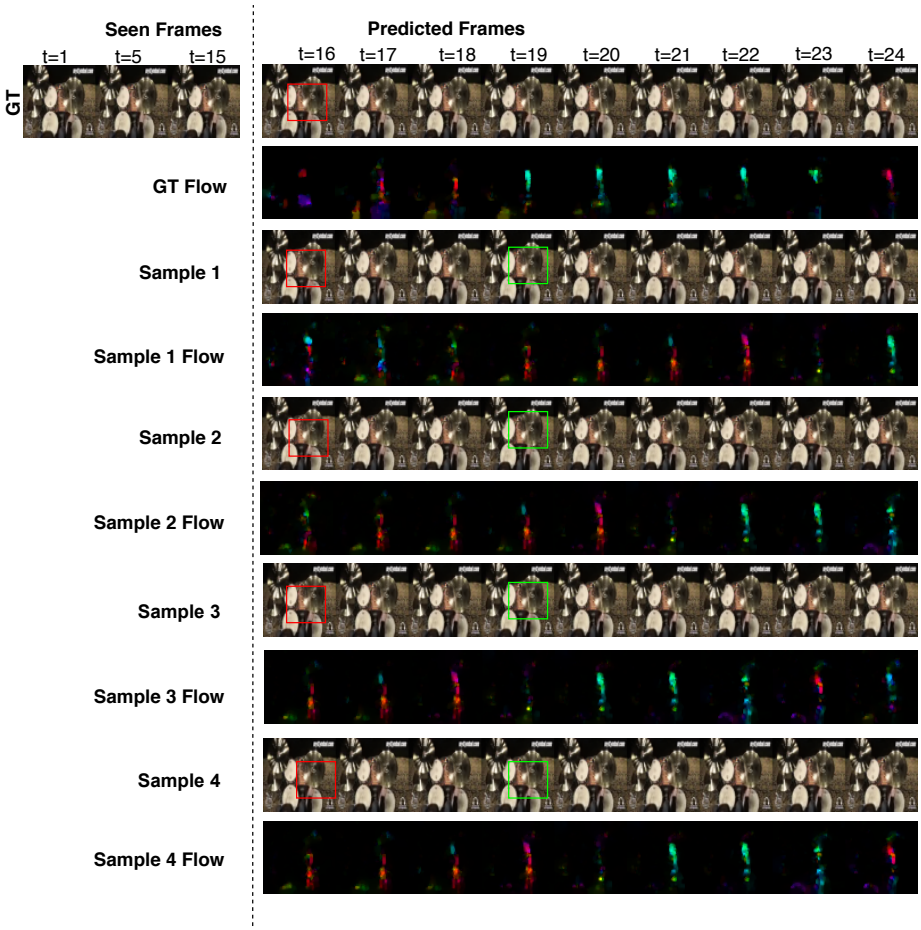


Fig. 92: Diverse sample generations on the AudioSet Drums dataset by our method along with the optical flows between frames. The red square denotes regions of high motion, while the green square highlights frames where noticeable differences are observed across samples.

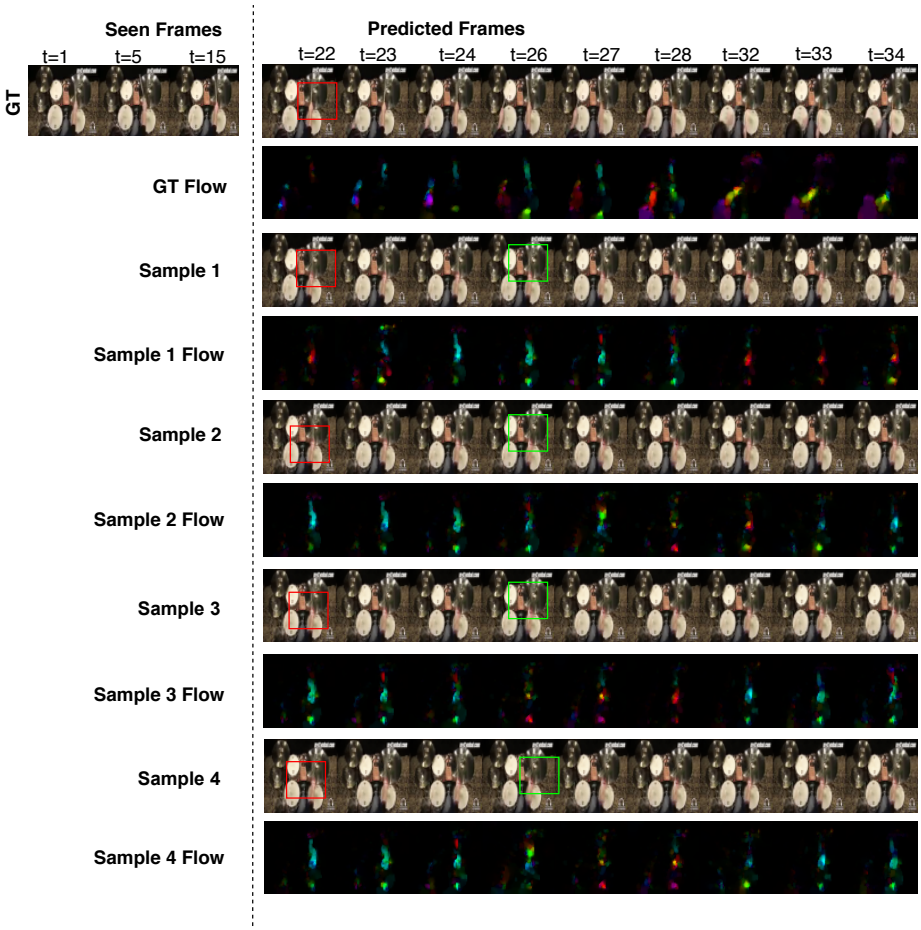


Fig. 93: Diverse sample generations on the AudioSet Drums dataset by our method along with the optical flows between frames. The red square denotes regions of high motion, while the green square highlights frames where noticeable differences are observed across samples.

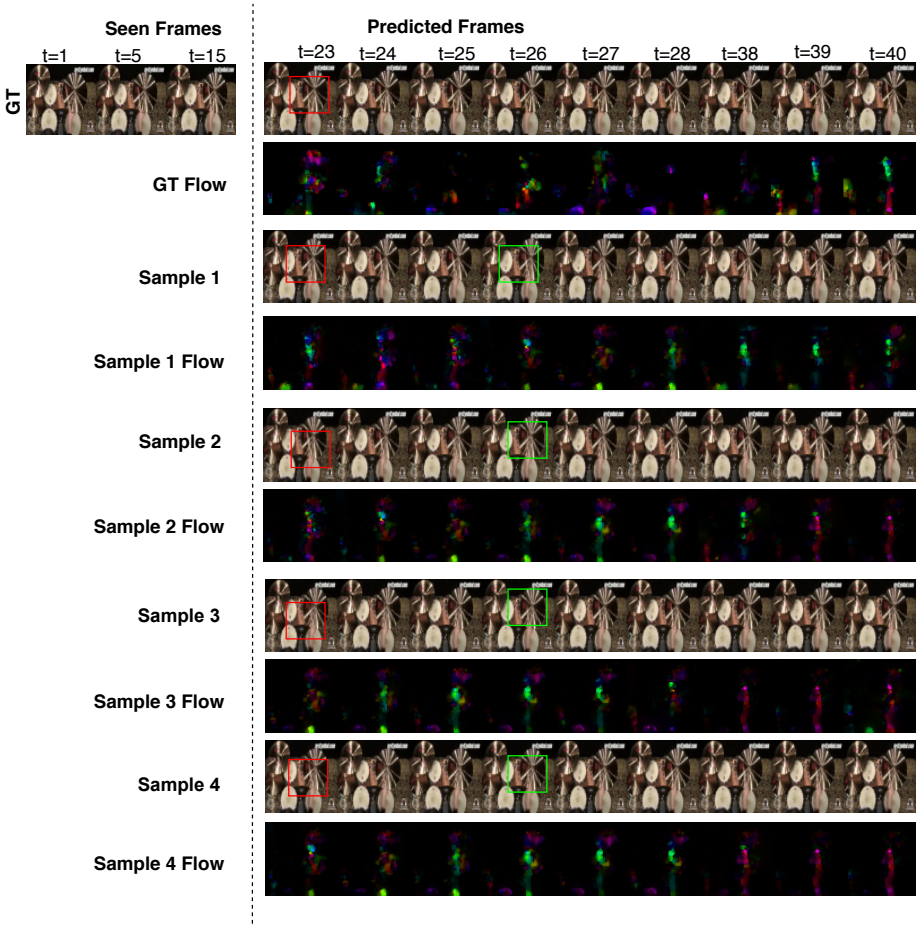


Fig. 94: Diverse sample generations on the AudioSet Drums dataset by our method along with the optical flows between frames. The red square denotes regions of high motion, while the green square highlights frames where noticeable differences are observed across samples.

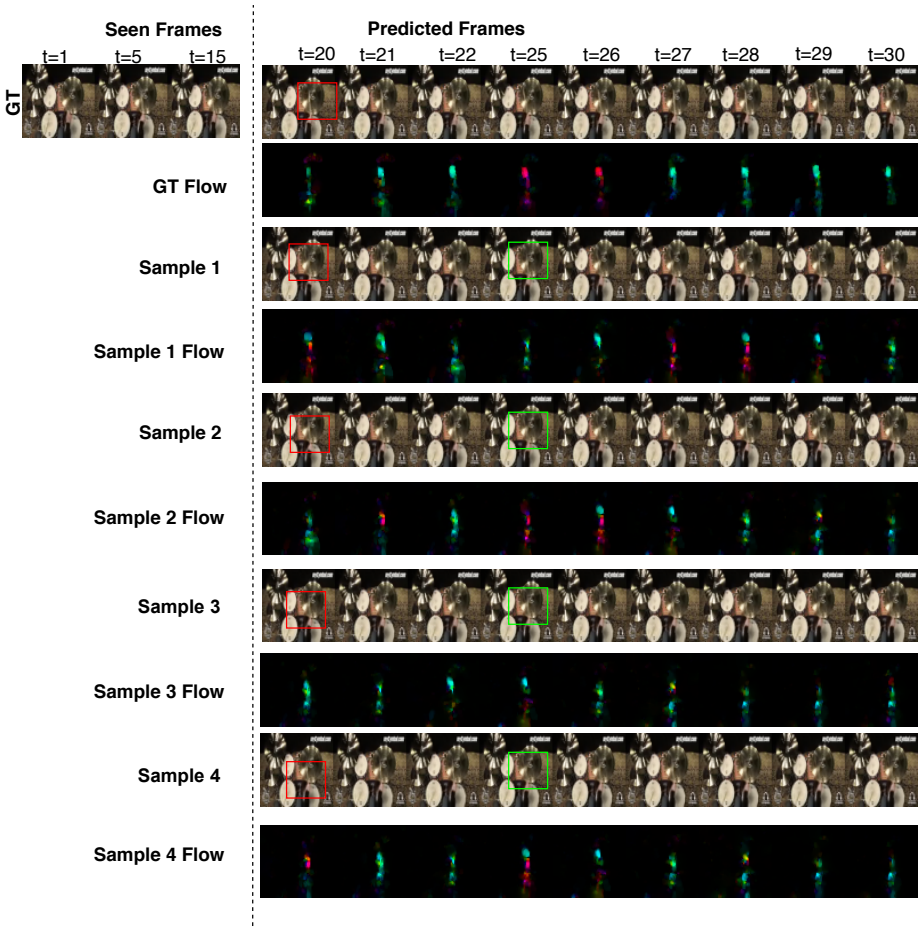


Fig. 95: Diverse sample generations on the AudioSet Drums dataset by our method along with the optical flows between frames. The red square denotes regions of high motion, while the green square highlights frames where noticeable differences are observed across samples.

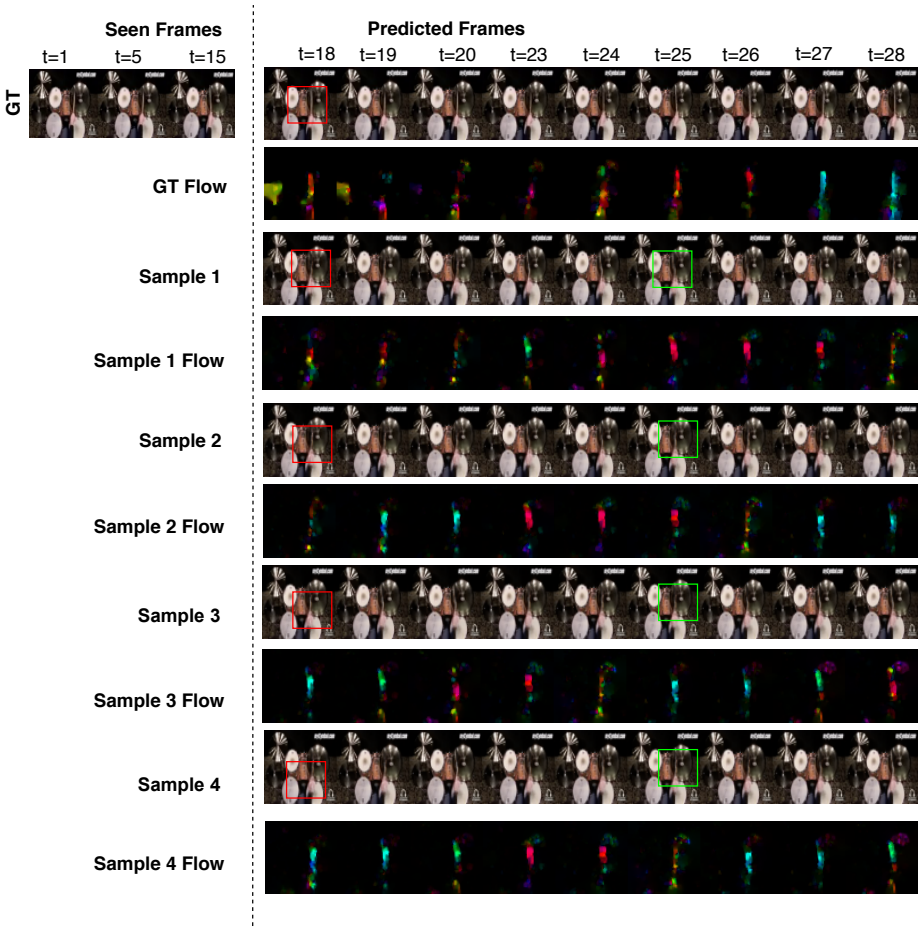


Fig. 96: Diverse sample generations on the AudioSet Drums dataset by our method along with the optical flows between frames. The red square denotes regions of high motion, while the green square highlights frames where noticeable differences are observed across samples.

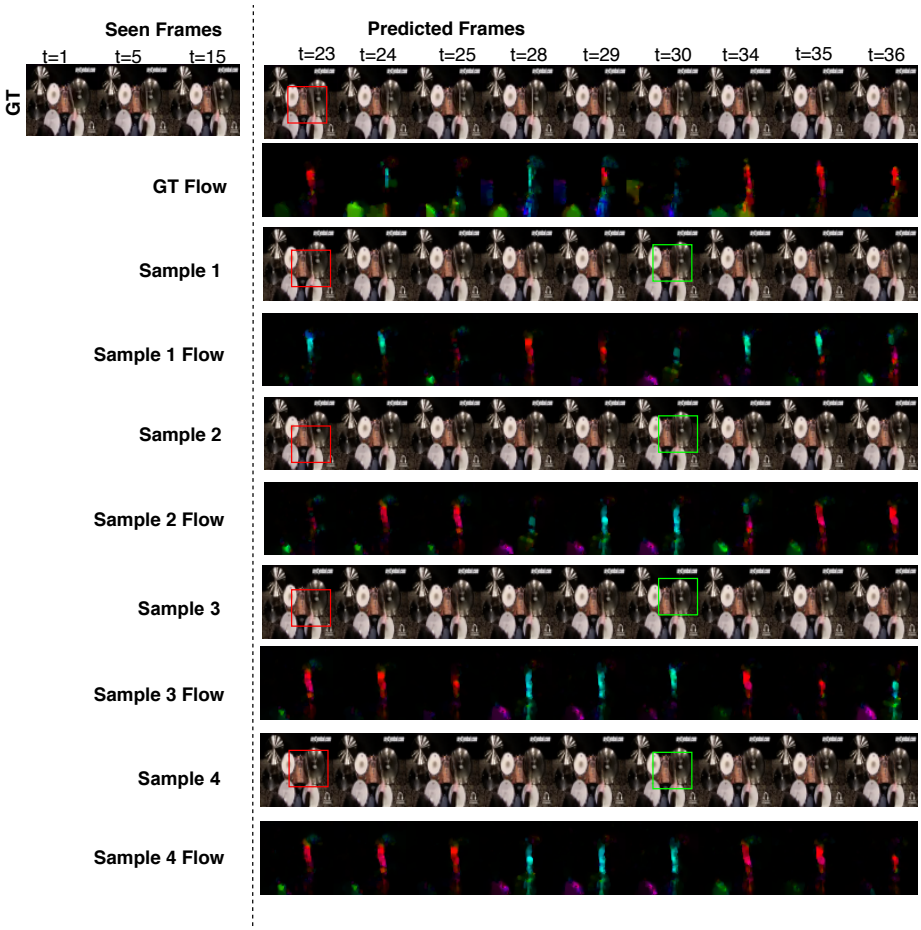


Fig. 97: Diverse sample generations on the AudioSet Drums dataset by our method along with the optical flows between frames. The red square denotes regions of high motion, while the green square highlights frames where noticeable differences are observed across samples.

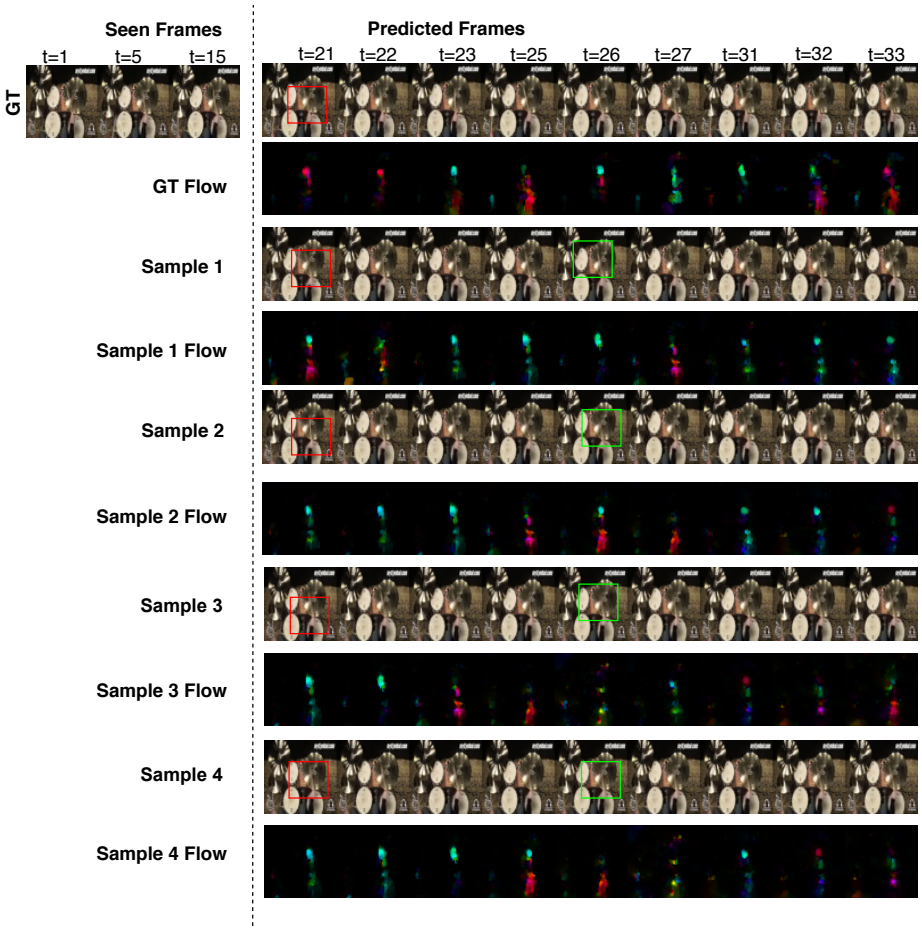


Fig. 98: Diverse sample generations on the AudioSet Drums dataset by our method along with the optical flows between frames. The red square denotes regions of high motion, while the green square highlights frames where noticeable differences are observed across samples.

7.2 Failure Cases

Figure 99 shows some scenarios where our model fails to generate visually compelling frames. This is mainly seen when the region of motion in the seen frames is localized to a small region. Our model, in such cases, essentially displays a static frame. This is typified by the slender '1' in Multimodal MovingMNIST or the limited hand motion (Figure 99) in the case of the YouTube Painting dataset. We intend to resolve this issue in our future work by replacing the Mean-Squared loss term in our objective, which uniformly penalizes all pixels, with a weighted version that would attend more to 'regions of interest' - where more motion is observed.

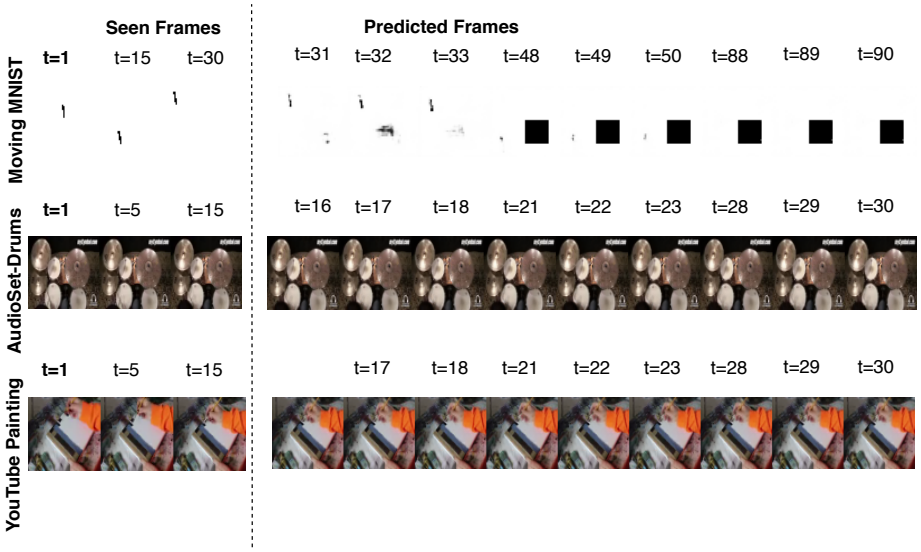


Fig. 99: An assortment of some of the failure cases of our method on the 3 datasets.

References

1. ASMR, T.: Painting ASMR (2019 (accessed November 5, 2019)), `https://www.youtube.com/playlist?list=PL5Y0dQ2DJHj47sK5jsbVkVpTQ9r7T090X`
2. Denton, E., Fergus, R.: Stochastic video generation with a learned prior. In: International Conference on Machine Learning. pp. 1182–1191 (2018)
3. Gemmeke, J.F., Ellis, D.P., Freedman, D., Jansen, A., Lawrence, W., Moore, R.C., Plakal, M., Ritter, M.: Audio set: An ontology and human-labeled dataset for audio events. In: 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 776–780. IEEE (2017)

4. Hsieh, J.T., Liu, B., Huang, D.A., Fei-Fei, L.F., Niebles, J.C.: Learning to decompose and disentangle representations for video prediction. In: Advances in Neural Information Processing Systems. pp. 517–526 (2018)
5. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
6. LeCun, Y., Bottou, L., Bengio, Y., Haffner, P., et al.: Gradient-based learning applied to document recognition. Proceedings of the IEEE **86**(11), 2278–2324 (1998)
7. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: Advances in neural information processing systems. pp. 5998–6008 (2017)
8. Vougioukas, K., Petridis, S., Pantic, M.: End-to-end speech-driven facial animation with temporal gans. arXiv preprint arXiv:1805.09313 (2018)
9. Yuan, X., Wang, T., Gulcehre, C., Sordoni, A., Bachman, P., Subramanian, S., Zhang, S., Trischler, A.: Machine comprehension by text-to-text neural question generation. arXiv preprint arXiv:1705.02012 (2017)