# A. Additional Details

## A.1. Network Architecture

Our spatial-temporal transformer comprises 20 spatial-temporal blocks. Each spatial-temporal block consists of a spatial layer followed by a temporal layer. The specific configurations are as follows:

- **Hidden States Dimension:** 1280
- **MLP Dimension in Attention Layers:** 5120
- **Number of Attention Heads:** 16
- **Diffusion Head:** Comprises 3 blocks with a dimension of 1024, implemented identically to the diffusion head in MAR [22].

Overall, our model consists of approximately **850** million parameters, which is on the same scale as the baseline models, ensuring a fair comparison.

## A.2. Data Sampling

### A.2.1. Training

During training, we uniformly sample a video from the training set and subsequently extract a clip based on a randomly selected frame interval. Specifically, with dynamic interval training, the frame interval is uniformly sampled from 1 to 25.

### A.2.2. FVD Evaluation

We use different sampling strategies for the UCF-101 and Kinetics-600 (K600) datasets for the real distribution:

- **UCF-101:** We randomly sample 2,048 videos and extract a single clip from each using a fixed frame interval of 3.
- **K600:** We randomly sample 50,000 videos and extract a single clip from each with a frame interval of 1.

We also generate a equal number of clips for FVD computation. It is important to note that MAGVIT-1 [51] utilizes a real distribution of 300,000 videos by applying 6 random spatial and temporal crops per video in the evaluation on K600. To maintain simplicity and ensure reproducibility, we limit our evaluation to 50,000 videos for the real distribution. Consequently, our Fréchet Video Distance (FVD) score does not benefit from the larger sample size used in MAGVIT-1.

## A.3. Training Procedure

The training process involves the following configurations:

- **Warmup Steps:** 10,000
- **Learning Rate:** $2 \times 10^{-4}$
- **Weight Decay:** 0.02
- **Input Frames:**
  - 17 frames when using a 3D-VAE
  - 16 frames when using a 2D-VAE
- **Batch Size:** 256 for K600 and 128 for UCF-101.
- **Training Epoch:** 150 for K600 and 1400 for UCF-101.
- **Patch Size in Input Layer:** 2 (consistent across all experiments)
- **Exponential Moving Average (EMA):** Applied with a decay rate of 0.9999. All reported results are generated using the EMA model.
- **Training Precision:** BF16, which is crucial for model convergence.

## A.4. Inference Procedure

During inference, the following settings are applied:

- **Default Masked Prediction Steps:** 64
- **Denoising Steps for Diffusion Head:** 100

It is important to note that:

- During training, each spatial layer employs fully bidirectional attention without any attention mask, while each temporal layer utilizes causal attention with a designed attention mask.
- At inference time, the attention mask is unnecessary as the model attends only to previously generated frames.
- When inferring longer sequences, we interpolate the temporal position embeddings using linear interpolation.
- In our proposed CTF paradigm, the sequence length of the model input is doubled solely during training. The sequence length remains unchanged during inference.

# B. Additional Analysis

## B.1. Class-conditioanl Video Generation

Tab. 4 shows MAGI's state-of-the-art performance in class-conditional video generation on UCF-101, leveraging classifier-free guidance[15] and the Cosmos VAE[26]. MAGI achieves an FVD of **76.1**, a significant improvement of approximately 60% over the recent OmniTokenizer baseline[45] (FVD 191). Additionally, MAGI surpasses the AR variant of MAGVIT-v2 [52] (FVD 109), setting a new standard for autoregressive video generation models in class-conditional settings. Notably, we maintain the same hyperparameters as those used in the unconditional counterpart.

Table 4. **Class-Conditional Video Generation Performance** on UCF-101. Evaluation protocol aligns with MAGVIT [51] $^\star$ indicates results of AR variants reported in MAGVIT-v2 [52]. $^\circ$ denotes zero-shot results from VideoPoet's paper. Please zoom in.

| Type | Method | $\text{FVD}_{128,\ 10k}$ |
|---|---|---|
| NAR | MAGVIT [51] | 76 |
| NAR | MAGVIT-v2 [52] | 58 |
| AR | MAGVIT [51] | 265$^\star$ |
| AR | MAGVIT-v2 [52] | 109$^\star$ |
| AR | VideoPoet [20] | 355$^\circ$ |
| AR | Omni [45] | 191 |
| AR | MAGI (Ours, CFG=7) | **76.1** |

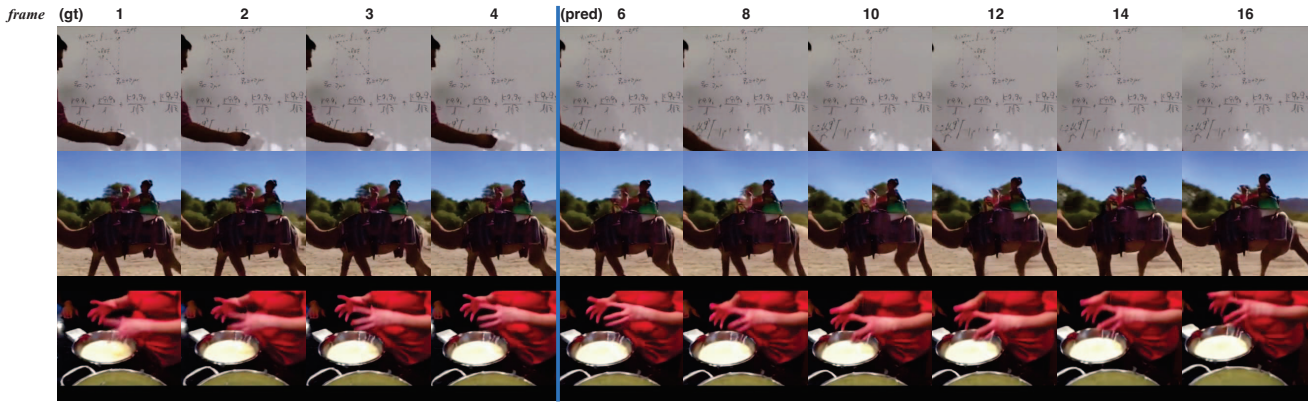


Figure 7. **Case Study: Video Prediction on Kinetics-600**. MAGI generates high-quality future frames conditioned on past frames.

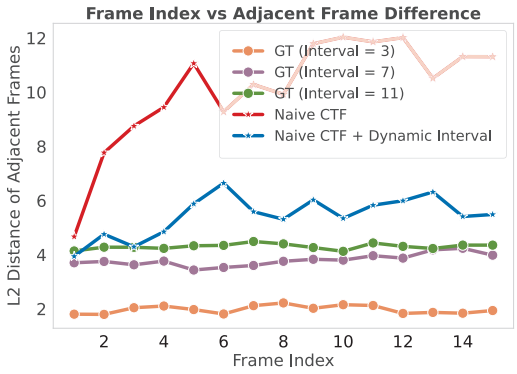


Figure 8. **Average L2 Distance** between adjacent frames on UCF-101 for 1,024 ground-truth and model-generated videos.

### B.2. Effectiveness of Dynamic Interval Training

We plot the L2 pixel distance between adjacent frames for CTF with and without dynamic interval training, along with ground truth (GT) sampled at varying frame intervals (see Figure 8). CTF without dynamic interval training exhibits abnormally increasing pixel differences due to error accumulation, while dynamic interval training significantly mitigates this issue, resulting in a more stable curve. Similar findings in HowFar [19] show that models trained on data with diverse motion strengths better adapt to intermediate motions. Thus, dynamic interval training enhances robustness and reduces exposure bias without additional data.

Since we compute FVD using GT frames at specific intervals (interval = 3 in UCF-101), it is necessary to include interval conditioning when using dynamic interval training. SDXL [27] also finds that conditioning embeddings help utilize data with diverse resolutions and aspect ratios, leading to better generation quality.

### B.3. Training Cost of CTF

While naive CTF doubles the token sequence length compared to MTF during training, increasing computational cost (4x), exploiting the sparsity in our attention mask with PyTorch's FlexAttention [¶] significantly improves efficiency by skipping computations in masked positions (Fig. 3 of the main paper). After optimization, CTF's training FLOPs are nearly identical to the naive implementation of MTF. Additionally, their inference FLOPs are the same.

## C. Further Analysis and Future Work

**MAGI: A Versatile and General Framework** Complete Teacher Forcing (CTF) offers a robust mechanism to convert bidirectional modeling, like MAR [22] or Diffusion [16] models, into efficient semi-autoregressive methods. This shift expands their applicability by combining bidirectional and autoregressive strengths. Theoretically, CTF and its training strategies are independent of innovations like the high-compression VAE in LTX-Video [12] or linear attention in SANA [49]. By integrating optimizations such as linear attention, model distillation, and high-quality datasets—we envision MAGI enabling real-time interactive video generation with outstanding performance.

**Future Research Directions** Future research can explore these ambitious directions:

- **Text-to-Video Generation:** Extending our method to text-to-video tasks while studying scalability in model size and datasets.
- **Interactive World Model:** Adapting the model to build a real-time interactive world model for dynamic, responsive environments.

[¶] https://pytorch.org/blog/flexattention/