Imagen Video

HIGH DEFINITION VIDEO GENERATION
WITH DIFFUSION MODELS

Presenter: Alex Chandler

02-29-2023

Motivation

- Recent widespread success in text-to-image diffusion models
  - Ex: Dalle2, Imagen
- Recent vast improvements in speed of diffusion and tricks for training stability
- Release of many other text-to-video papers in 2022
  - Ex: Ho et al., 2022b, Singer et al. (2022)

“An impressionist portrait, by Claude Monet, 1875, of Buffalo Bayou with a view of the Houston downtown skyline.” - Dalle2

Challenges

- What makes this such a difficult problem to solve?
  - Need for temporal accuracy and temporal coherence
  - Computational Requirements
  - Greater understanding of language and natural environments
  - Controllability
  - Generation Diversity
  - 3D Object Understanding
  - Handling artistic styles

Previous Work

Diffusion Models:

- Important Papers: Sohl-Dickstein et al., 2015; Song & Ermon,2019; Ho et al., 2020
- Cascaded Diffusion: Ho et al., 2022a

Text-to-Image:

- DALL-E 2: Ramesh et al., 2022, Imagen: Saharia et al., 2022b)

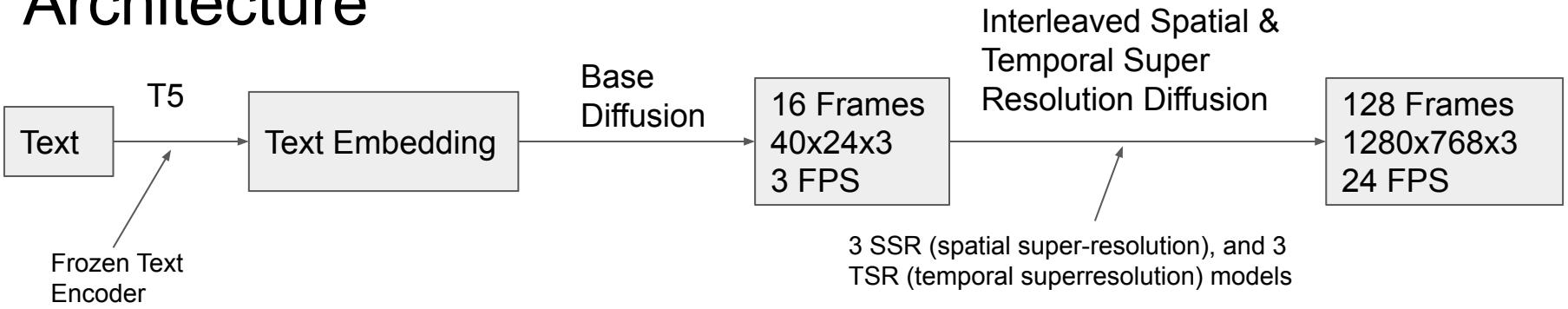
Text-to-Video:

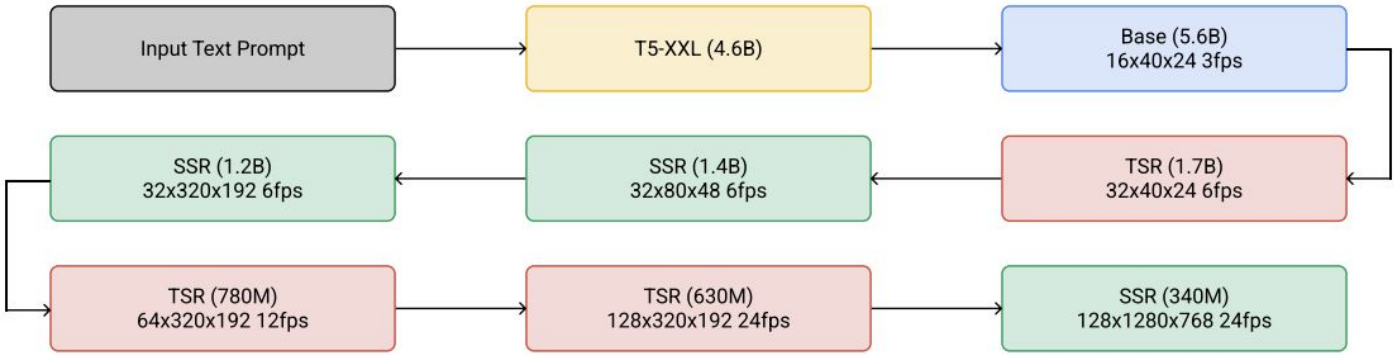
- Video Diffusion Models: Ho et al., 2022b (From Google)
- Make-A-Video: Singer et al. 2022 (From Meta)

Other:

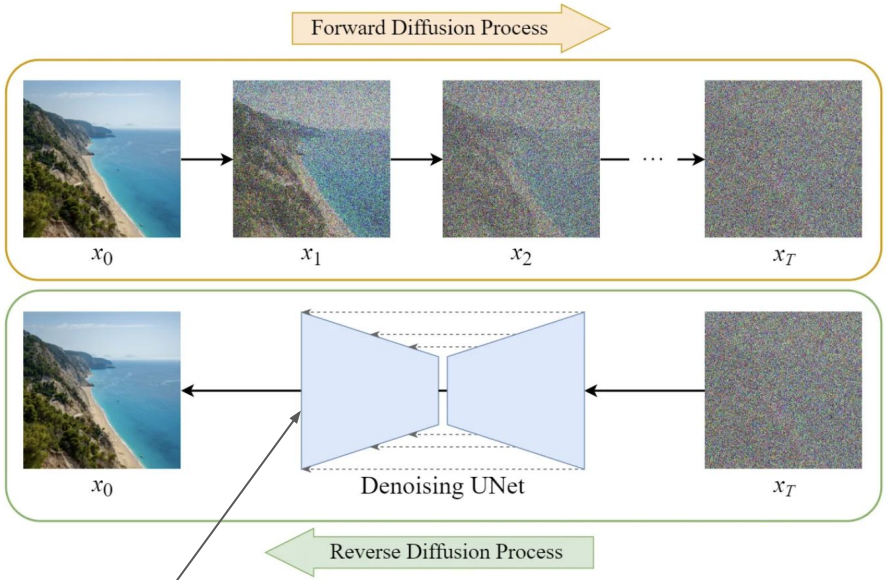
- Progressive Distillation: Salimans & Ho, 2022
- U-Net architecture
- Classifier Guidance vs Classifier-Free Guidance
- v-parametrization: Salimans & Ho, 2022
- And so much more!

Architecture



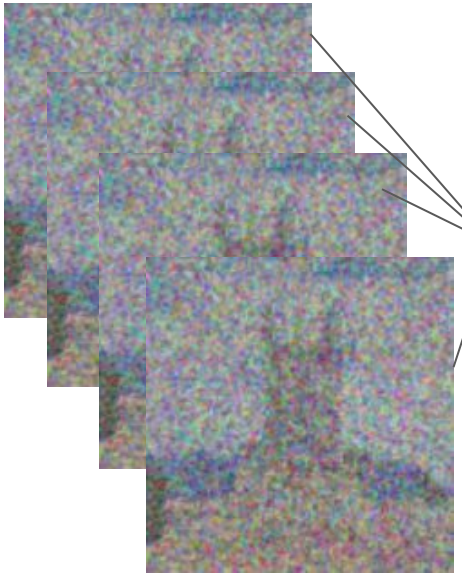


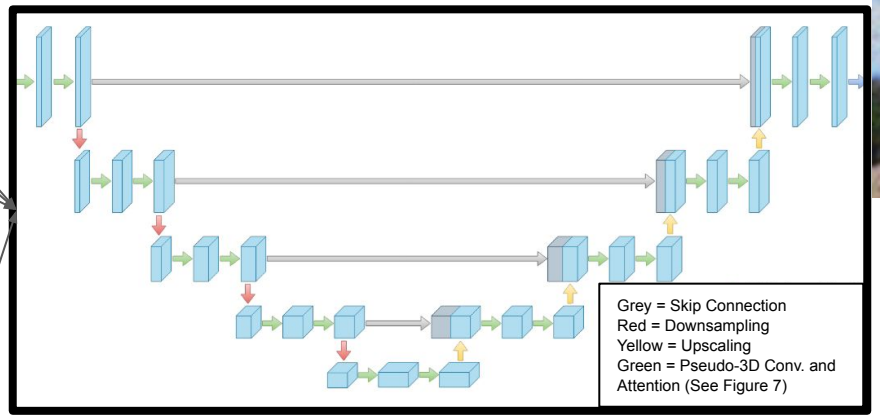
Base Image Diffusion Model



Can we extend an image diffusion model for video?

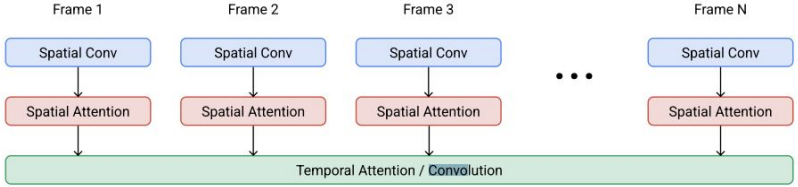
Base Video Diffusion Model



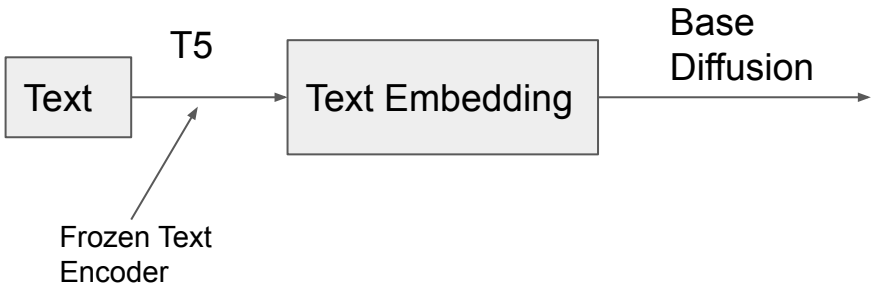


= 16 frame video at 40×24 resolution and 3 frames per second

Figure 7



Base Video Diffusion Model



16 frame video at 40×24 resolution and 3 frames per second

Cascade Diffusion

1. Initially generate an image or video at a low resolution
2. Sequentially increase the resolution of the image or video through a series of super-resolution diffusion models.
3. Text embeddings are injected into all models, not just the base model

Why?

- Can model problems of high dimensionality while keeping each sub-model relatively simple.
- Reduces Computation
    - Each diffusion model is trained independently, allowing for parallel training of all 7 models

Fully-Convolutional Temporal and Spatial Super-Resolution Models

Training:

Goal: Achieve desired spatial resolution and more frames and temporal resolutions:

Method: Spatial resizing and frame skipping

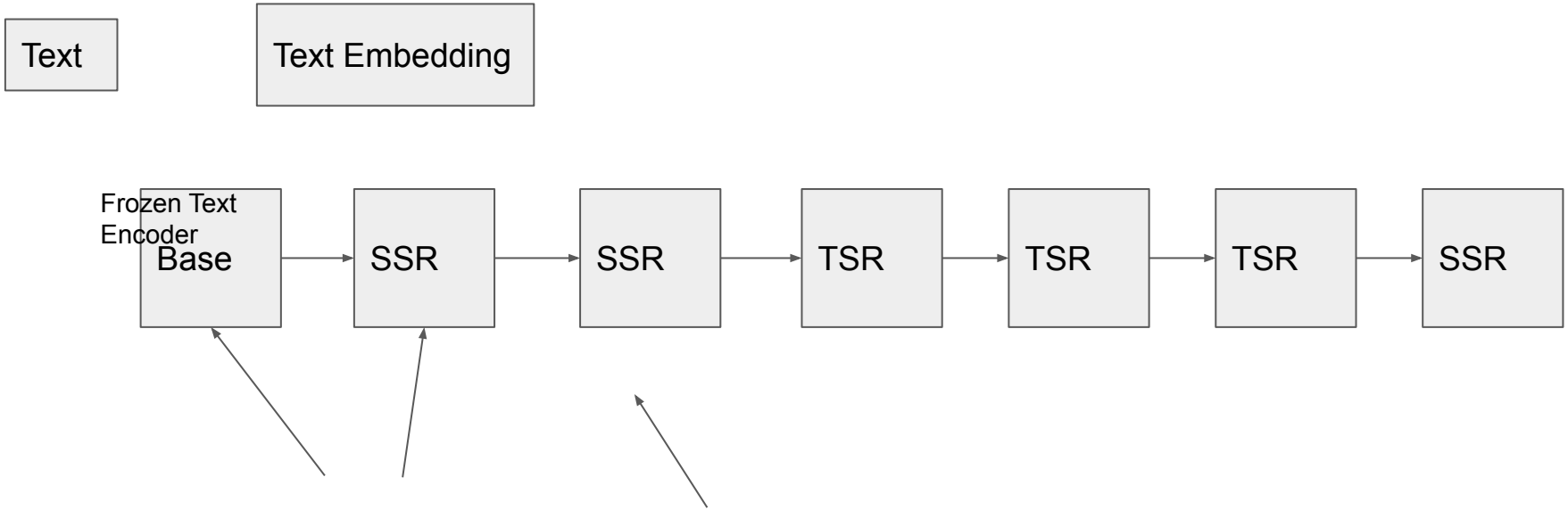
Generation Time:

- SSR models increase spatial resolution for all input frames
- TSR models increase temporal resolution by filling in intermediate frames between input frame
- All models generate an entire block of frames simultaneously

Cascade Diffusion Training

- Each diffusion model is trained independently
  - Allows for parallel computation
- Models trained on both images and video
  - Images treated as single video frames
    - Pack independent images into sequences as video, but bypass the temporal convolutional residual blocks by masking out their computation path. Similarly, disable cross-frame temporal attention by applying masking to the temporal attention maps.
- Last SSR model trained on random lower resolution spatial crops
- A lot of other tricks
  - “Following Ho et al. (2022a), we apply Gaussian noise augmentation with a random signal-to-noise ratio to the conditioning input video during training.” (Same as Imagen)
  - Classifier Free Guidance (Same as Imagen)
  - Dynamic Thresholding (Same as Imagen)
  - Use deterministic DDIM sampler (Same as Imagen)

Cascade Diffusion Training



$$\mathcal{L}(\mathbf{x}) = \mathbb{E}_{\boldsymbol{\epsilon}\sim\mathcal{N}(0,\mathbf{I}),t\sim U(0,1)}\left[\|\hat{\boldsymbol{\epsilon}}_\theta(\mathbf{z}_t,\lambda_t) - \boldsymbol{\epsilon}\|_2^2\right]$$

$$\mathbf{z}_t = \alpha_t\mathbf{x} + \sigma_t\boldsymbol{\epsilon}, \text{ and } \hat{\boldsymbol{\epsilon}}_\theta(\mathbf{z}_t,\lambda_t) = \sigma_t^{-1}(\mathbf{z}_t - \alpha_t\hat{\mathbf{x}}_\theta(\mathbf{z}_t,\lambda_t))$$

Parameterize models in terms of the v-parameterization (Salimans & Ho, 2022)

rather than predicting $\boldsymbol{\epsilon}$ or $\mathbf{x}$ directly;

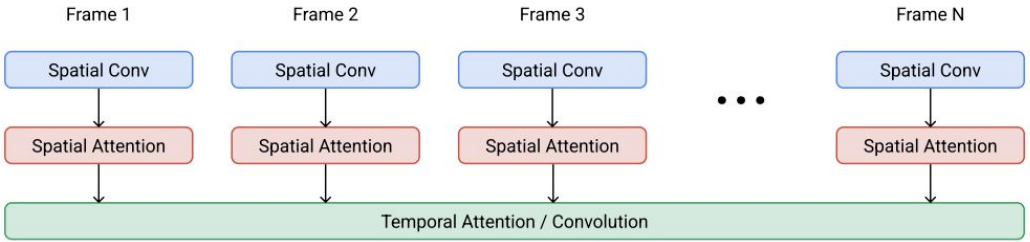
Use of discrete time ancestral sampler

Starting at $\mathbf{z}_1 \sim \mathcal{N}(\mathbf{0},\mathbf{I})$, the ancestral sampler follows the rule

$$\mathbf{z}_s = \tilde{\boldsymbol{\mu}}_{s|t}(\mathbf{z}_t,\hat{\mathbf{x}}_\theta(\mathbf{z}_t)) + \sqrt{(\tilde{\sigma}_{s|t}^2)^{1-\gamma}(\sigma_{t|s}^2)^{\gamma}}\boldsymbol{\epsilon}$$

Video U-Net

- U-NET enhanced with temporal attention and temporal convolution
- Combine individual images together which are a part of the same video
- Trained on images and videos simultaneously
  - Single images are treated as single frame videos



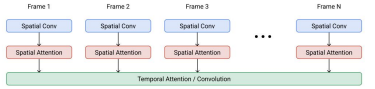


Figure 7: Video U-Net space-time separable block. Spatial operations are performed independently over frames with shared parameters, whereas the temporal operation mixes activations over frames. Our base model uses spatial convolutions, spatial self-attention and temporal self-attention. For memory efficiency, our spatial and temporal super-resolution models use temporal convolutions instead of attention, and our models at the highest spatial resolution do not have spatial attention.

Video U-Net

- U-NET enhanced with temporal attention and temporal convolution
- Combine individual images together which are a part of the same video
- Trained on images and videos simultaneously
  - single images are treated as single frame videos

Diffusion (Basic)

Classifier Guidance: Dhariwal & Nichol (2021)

First what is Classifier Guidance?

- Boosts diffusion quality with an extra trained classifier
- Mixes a diffusion model's score estimate with the input gradient of the logged probability of a classifier.
- Enables diffusion model to condition on a trained classifier labels, guiding a sample towards the desired label with the gradients from the classifier
- Can be used as a hyperparameter to specify how closely to follow prompt
    - Trade-off of Diversity vs Quality

Classifier Free Guidance: Ho & Salimans (2022)

Key Idea:

- Mix the score estimates of a conditional diffusion model and a jointly trained unconditional diffusion model
- Interpolates between predictions from a diffusion model with and without labels.
- Does not require training of any separate classifying model.
- Note: details are really complicated

Benefits

- Does not rely on knowledge from separately trained, and usually smaller, classification model.
- More robust gradients
- Better empirical results

Progressive Distillation

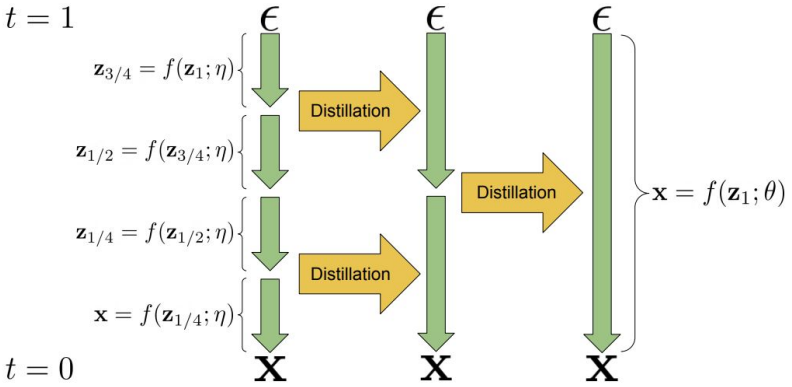
- Introduced in separate Google Research, Brain Team 2022 Paper called Progressive Distillation For Fast Sampling.
- Purpose: Reduce number of diffusion steps by teacher-student knowledge distillation
- The authors of the paper could reduce the number of steps from 8192 steps to 4. (204,800% speed increase).

Progressive Distillation in a Nutshell

1. Train a teacher diffusion model (It will be slow and may take up to 8,192 steps, but lead to high quality input)
2. Create new model called student
    a. Student is a clone with the same weights and parameter as teacher
3. Denoise image by running noisy image twice through teacher model
    a. (X->X') then (X'->X'')
4. Feed student the original noisy images (X) and expect it to output two iterations of denoising (X'').
    a. Effectively trains student model to be twice as fast as the the teacher model
5. Now set the teacher model to the new faster student model
6. Repeat steps 2-5

Progressive Distillation Old Version

Progressive Distillation:
Introduced in separate Google Research, Brain Team 2022 Paper called Progressive Distillation For Fast Sampling. Main idea of this paper is a novel way to reduce the number of diffusion steps, up to low as four, without losing much quality, through what they call progressive distillation.



Algorithm 1 Standard diffusion training

Require: Model $\hat{\mathbf{x}}_\theta(\mathbf{z}_t)$ to be trained
Require: Data set $\mathcal{D}$
Require: Loss weight function $w()$

while not converged **do**
- $\mathbf{x} \sim \mathcal{D}$ ▷ Sample data
- $t \sim U[0, 1]$ ▷ Sample time
- $\epsilon \sim N(0, I)$ ▷ Sample noise
- $\mathbf{z}_t = \alpha_t \mathbf{x} + \sigma_t \epsilon$ ▷ Add noise to data
- $\tilde{\mathbf{x}} = \mathbf{x}$ ▷ Clean data is target for $\hat{\mathbf{x}}$
- $\lambda_t = \log[\alpha_t^2 / \sigma_t^2]$ ▷ log-SNR
- $L_\theta = w(\lambda_t) \|\tilde{\mathbf{x}} - \hat{\mathbf{x}}_\theta(\mathbf{z}_t)\|_2^2$ ▷ Loss
- $\theta \leftarrow \theta - \gamma \nabla_\theta L_\theta$ ▷ Optimization

end while

Algorithm 2 Progressive distillation

Require: Trained teacher model $\hat{\mathbf{x}}_\eta(\mathbf{z}_t)$
Require: Data set $\mathcal{D}$
Require: Loss weight function $w()$
Require: Student sampling steps N

for K iterations **do**
- $\theta \leftarrow \eta$ ▷ Init student from teacher
- **while** not converged **do**
  - $\mathbf{x} \sim \mathcal{D}$
  - $t = i/N, \; i \sim Cat[1, 2, \ldots, N]$
  - $\epsilon \sim N(0, I)$
  - $\mathbf{z}_t = \alpha_t \mathbf{x} + \sigma_t \epsilon$
  - `# 2 steps of DDIM with teacher`
  - $t' = t - 0.5/N, \quad t'' = t - 1/N$
  - $\mathbf{z}_{t'} = \alpha_{t'} \hat{\mathbf{x}}_\eta(\mathbf{z}_t) + \frac{\sigma_{t'}}{\sigma_t}(\mathbf{z}_t - \alpha_t \hat{\mathbf{x}}_\eta(\mathbf{z}_t))$
  - $\mathbf{z}_{t''} = \alpha_{t''} \hat{\mathbf{x}}_\eta(\mathbf{z}_{t'}) + \frac{\sigma_{t''}}{\sigma_{t'}}(\mathbf{z}_{t'} - \alpha_{t'} \hat{\mathbf{x}}_\eta(\mathbf{z}_{t'}))$
  - $\tilde{\mathbf{x}} = \frac{\mathbf{z}_{t''} - (\sigma_{t''}/\sigma_t)\mathbf{z}_t}{\alpha_{t''} - (\sigma_{t''}/\sigma_t)\alpha_t}$ ▷ Teacher $\hat{\mathbf{x}}$ target
  - $\lambda_t = \log[\alpha_t^2 / \sigma_t^2]$
  - $L_\theta = w(\lambda_t) \|\tilde{\mathbf{x}} - \hat{\mathbf{x}}_\theta(\mathbf{z}_t)\|_2^2$
  - $\theta \leftarrow \theta - \gamma \nabla_\theta L_\theta$
- **end while**
- $\eta \leftarrow \theta$ ▷ Student becomes next teacher
- $N \leftarrow N/2$ ▷ Halve number of sampling steps

end for

V-Parameterization for Progressive Distillation

V-Parameterization Alleged Benefits:

- Avoids color shifting artifacts that are known to affect high resolution diffusion models
- Avoids temporal shifting
- Faster convergence of sample quality metrics
- Helpful for progressive distillation

v-prediction parameterization ($\mathbf{v}_t \equiv \alpha_t \boldsymbol{\epsilon} - \sigma_t \mathbf{x}$)

$\alpha_t \boldsymbol{\epsilon}$ = our sample noise

$\sigma_t \mathbf{x}$ = our sample data

where $\mathbf{z}_t = \alpha_t \mathbf{x} + \sigma_t \boldsymbol{\epsilon}$, and $\hat{\boldsymbol{\epsilon}}_\theta(\mathbf{z}_t, \lambda_t) = \sigma_t^{-1}(\mathbf{z}_t - \alpha_t \hat{\mathbf{x}}_\theta(\mathbf{z}_t, \lambda_t))$. We will drop the dependence on λ_t to simplify notation. In practice, we parameterize our models in terms of the $\mathbf{v}$-parameterization (Salimans & Ho, 2022), rather than predicting $\boldsymbol{\epsilon}$ or $\mathbf{x}$ directly; see Section 2.4.

Purpose of predicting v: allows us to predict $\bar{x}$

Predicting $\mathbf{v} \equiv \alpha_t \epsilon - \sigma_t \mathbf{x}$, which gives $\hat{\mathbf{x}} = \alpha_t \mathbf{z}_t - \sigma_t \hat{\mathbf{v}}_\theta(\mathbf{z}_t)$

V-parameterization

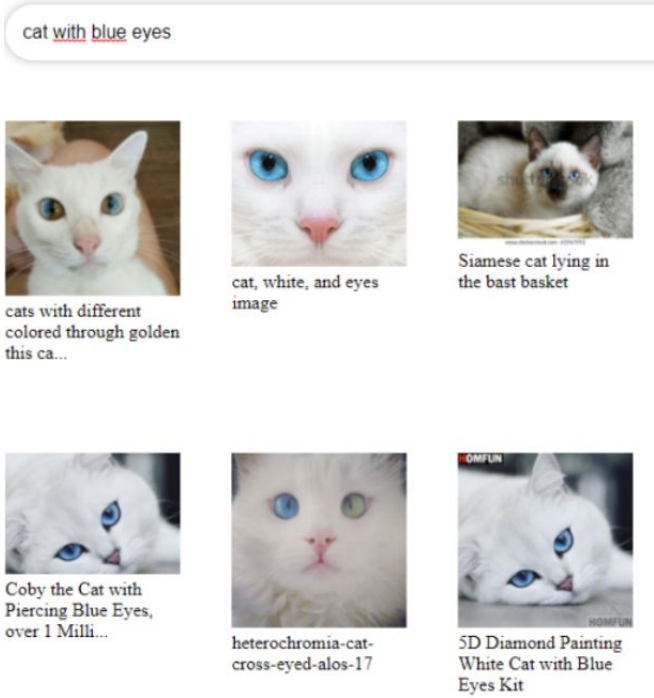
Distillation

- Goal: Quicken the speed of diffusion
- Definition:
- Details:

Dataset

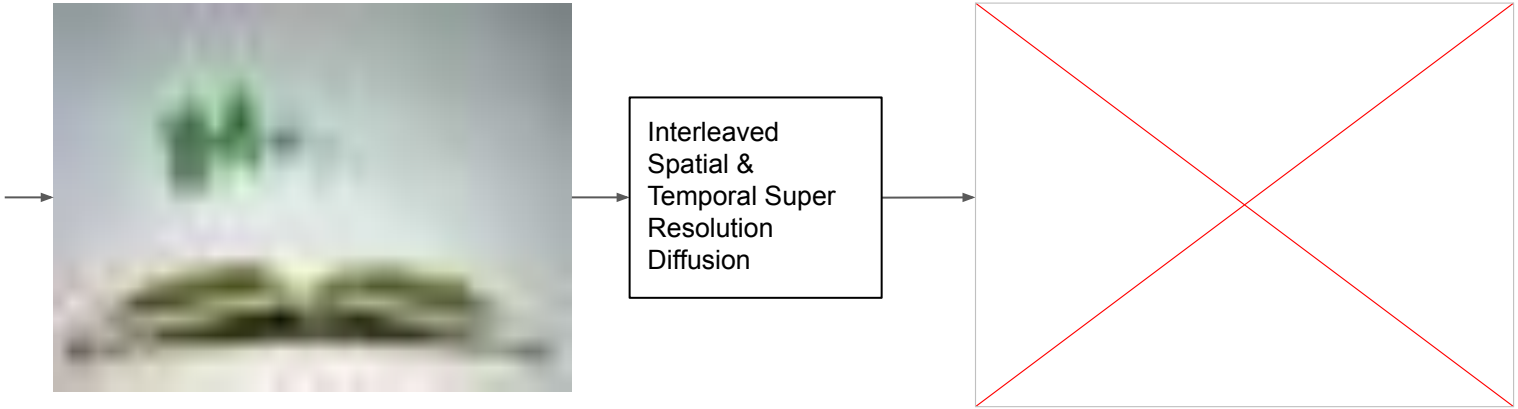
- Trained on a combination of a private internal dataset consisting of 14 million video-text pairs and 60 million image-text pairs.
- Also trained on the publicly available LAION-400M image-text dataset
- Information about LAION-400M
  - Freely-accessible dataset of image-text pairs filtered by CLIP with cosine similarity >=.3

LAION-400M Sample Query Images



Example 1

“Sprouts in the shape of text 'Imagen' coming out of a fairytale book”



Example 2

“A happy Elephant Wearing a birthday hat walking under the sea.”

A Whole Bunch of More Examples

1. “A cat on the left of a dog.”
2. “A teddy bear running in New York City.”
3. “A sheep to the right of a wine glass.”
4. “A tiny sprout coming out of the ground.”

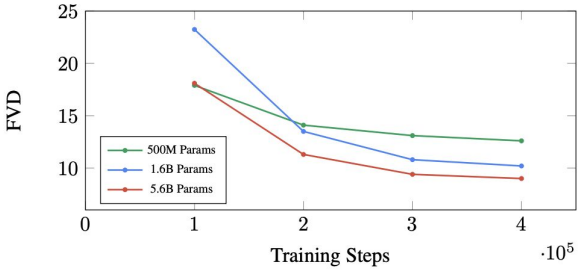
Evaluation

Evaluation

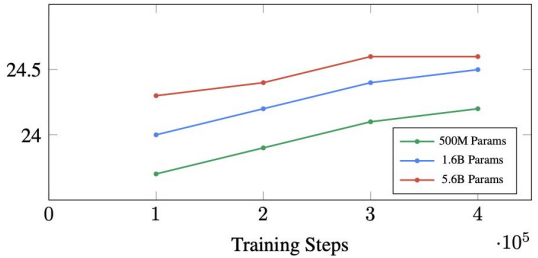
- Fréchet inception distance (FID) on individual frames (Heusel et al., 2017)
  - assess the quality of images created by a generative model
  - measures the distance between the feature representations of the generated images and the real images
- Fréchet Video Distance (FVD) (Unterthiner et al., 2019) for temporal consistency
- Frame-wise CLIP scores (Hessel et al., 2021) for video-text alignment
  - CLIP score calculation:
    - take the average score over all frames.
  - For CLIP R-Precision (Park et al., 2021):
    - compute the top-1 accuracy (i.e. R = 1)

Scaling Experiment

- Strongly benefits from scaling up the parameter count of the video U-Net.
  - Scaling only by increasing the base channel count and depth of the network



(a) Scaling Comparison on FVD scores.



(b) Scaling Comparison on CLIP scores.

Perceptual Quality and Distillation Experiment

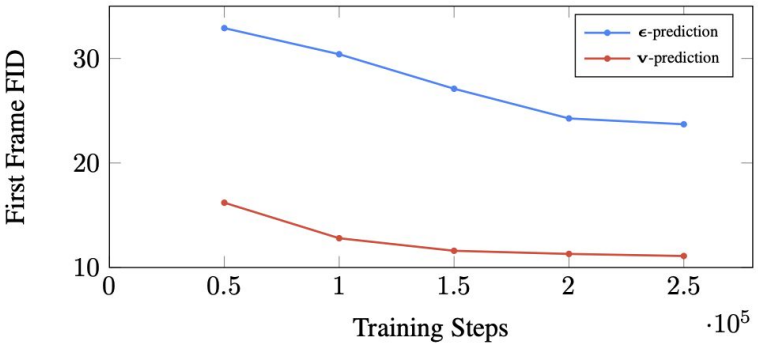
- Perceptual Quality and Distillation
  - Measured CLIP score and CLIP R-Precision
  - Distilled Models do not have noticeable drop in CLIP metrics.

Guidance w	Base Steps	SR Steps	CLIP Score	CLIP R-Precision	Sampling Time
constant=6	256	128	25.19±.03	92.12±.53	618 sec
oscillate(15,1)	256	128	25.02±.08	89.91±.96	618 sec
constant=6	256	8	25.29±.05	90.88±.50	135 sec
oscillate(15,1)	256	8	25.15±.09	88.78±.69	135 sec
constant=6	8	8	25.03±.05	89.68±.38	35 sec
oscillate(15,1)	8	8	25.12±.07	90.97±.46	35 sec
ground truth			24.27	86.18	

Table 1: CLIP scores and CLIP R-Precision (Park et al., 2021) values for generated samples and ground truth videos on prompts from our test set. Cells highlighted in green represent distilled

Other “Experiment”

- “Unique Video Generation Capabilities”
  - “capable of generating videos with artistic styles”
  - “3D consistency over the course of rotation is not exact”
  - “Imagen Video shows that video models can serve as effective priors for methods that do force 3D consistency”
- Comparing Prediction Parameterizations
  - ϵ-prediction models performed worse than v-prediction, especially at high resolutions
  - “ϵ-prediction prediction converges relatively slowly in terms of sample quality metrics and suffers from color shift and color inconsistency across frames in the generated videos”
  - v-parameterization converges faster than ϵ-prediction



Critique: Google Internal Text-Video Dataset

Strengths

- Unknown!
- 14 Million video-text pairs should be large enough for generability
  - Of course this depends on the diversity and details of internal dataset

Limitations:

- Google Internal Text-Video dataset not open to public
- Could it be youtube videos?
- What biases exist from the dataset?
- What filtration has been done on the dataset?

Critique: LAION-400M Image-Text Dataset:

Strengths

- Large dataset
- Some level of filtering (CLIP-filtered 400 million image-text pairs)
- Shown strong empirical results in existing SOTA text-image models (ex: Imagen)

Limitations:

- Includes a wide range of inappropriate content (racism, damaging social stereotypes, pornographic material): Birhane et al., 2021

Critique: Model Architecture

- Strengths
  - Reduction in computation with…:
    - Pseudo-3D Convolution
    - Cascade Diffusion
    - Progressive Distillation
    - V-Parameterization
    - Parallel Training
    - etc.
- Limitations
  - Lack of insight for why things work
  - Computation requirements
    - 4.6 Billion Parameters for T5 XXL
    - 11.6B diffusion model parameters

Critique: Model Architecture (Continued)

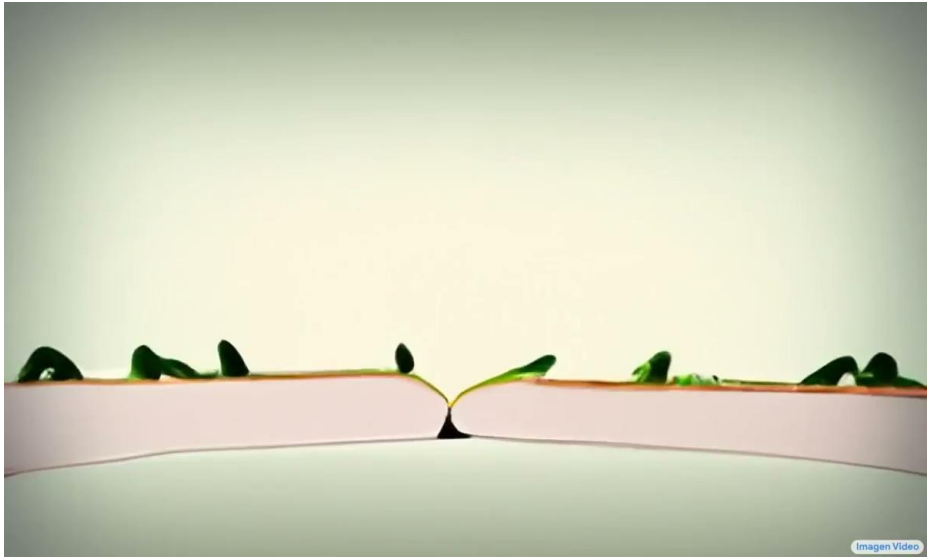
- Strengths
  - v-parameterization for numerical stability
    - avoids color shifting artifacts
    - faster convergence
  - Video-Image Joint Training
    - Authors found increase video quality
- Limitations
  - Can only generate short videos (increasing attention length quadratically increases computation)
  - Sub-optimal methods to ensure temporal coherence

Critique: Evaluation

- Strengths
  - Well, at least they tried
    - Attempts to measure image quality with FID and temporal coherence with FVD
    - Attempts to measure text-video alignment with CLIP and CLIP-R
- Limitations
  - Lacking real benchmark
    - Not evaluating ability to handle OOD prompts, rare words, misspellings, etc.
  - Lacking evaluation in comparison to Imagen (and other text-to-image evaluation)
  - Not measuring generation diversity
  - CLIP score calculation treats each frame equally by averaging score over all frames
  - And many more …

Summary

- Impressive, but still a bunch of issues
  - Waiting for Imagen Video 2
- Surprising rate of improvement in field
- Explosion of work in the field
- Research originating only from a few companies (Google, Meta, …)
- Lack of Understandability for design choices
  - Especially true with Imagen Video



What I would have done differently

- Use both T5 and CLIP
  - We desire grounding, and perhaps combining both T5-XXL and CLIP would lead to better results
- More for understandability
  - Visualization on the training structure of the model
  - More ablation studies would likely shown what tricks are most important
- More information about Google Internal Dataset
  - include limitations and biases
- Add more ablation studies
  - Word “ablation” not mentioned in paper
- Generate longer videos with Shifted Window Based Self-Attention (SWIN like)

Future Work for Paper / Reading

- Actual Current Work:
  - Not much (published **yet**)! Only came out four months ago
  - Concurrent work from Google Phenaki - “Realistic video generation from open-domain textual descriptions"
- Planned Future Work
  - Explore hybrid pipelines of multiple model classes further in future work

Extended Readings

Good Sources for Explainability:

- **Blogs:**
  - DDIM
  - Classifier Free Diffusion Guidance
- **Papers**:
  - DDPM
  - Classifier Free Diffusion Guidance
  - Progressive Distillation for Fast Sampling of Diffusion Models

Concurrent Work:

- Make-A-Video: Text-to-Video Generation without Text-Video Data
- Phenaki: Variable Length Video Generation from Open Domain Textual Descriptions

Thank You