Imagen Video: High Definition Video Generation with Diffusion Models

LLVM Paper Discussion

10.23.23

Bug Hunter

Alexandre Kaiser - amk1004

The role of a Bug Hunter

- Adversarial reader
- Find issues with
  - Rigor
  - Correctness
  - Reproducibility
  - Clarity
- Challenge the choices that were made

The role of a Bug Hunter

- Adversarial reader
- Find issues with
  - Rigor
  - Correctness
  - Reproducibility
  - Clarity
- Challenge the choices that were made

FID

FVD

CLIP scores

The role of a Bug Hunter

- Adversarial reader
- Find issues with
  - Rigor
  - Correctness
  - Reproducibility
  - Clarity
- Challenge the choices that were made

FID

FVD

CLIP scores

Correct math

The remainder is only empirically driven

The role of a Bug Hunter

- Adversarial reader
- Find issues with
  - Rigor
  - Correctness
  - Reproducibility
  - Clarity
- Challenge the choices that were made

FID

FVD

CLIP scores

Correct math

The remainder is only empirically driven

Not public

Proprietary dataset

The role of a Bug Hunter

- Adversarial reader
- Find issues with
  - Rigor
  - Correctness
  - Reproducibility
  - Clarity
- Challenge the choices that were made

FID

FVD

CLIP scores

Correct math

The remainder is only empirically driven

Not public

Proprietary dataset

Clear if and only if you read their previous work

FID

FVD

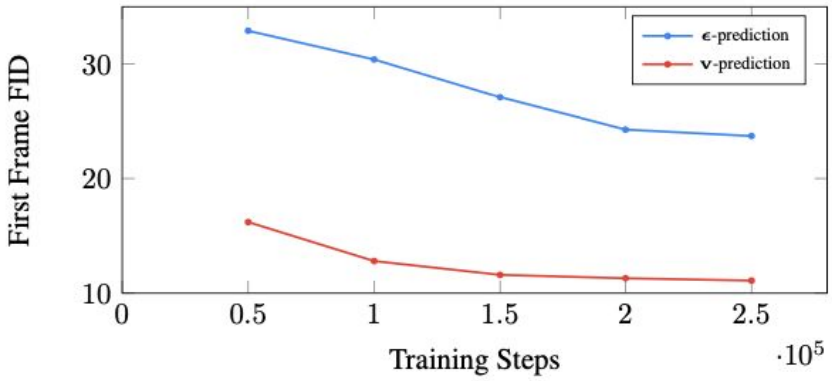
- Not an absolute metric
- Interpretable reporting must include comparison to other model performances

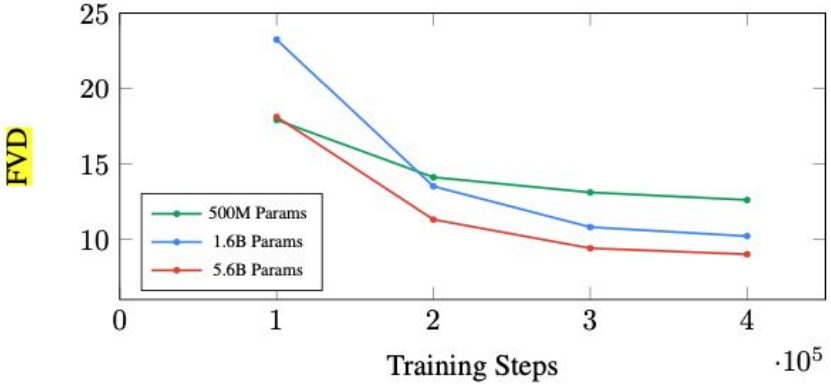
we evaluated Imagen Video on several different metrics, such as FID FVD and frame-wise CLIP scores for video-text alignment. Below, we explore 1) *scaling* 2) *parameterization* 3) *distilling*

FID

FVD

- Not an absolute metric
- Interpretable reporting must include comparison to other model performances





Evaluates ε-prediction vs v-prediction

Evaluates parameter size of Base model

NYU

CLIP scores

- **Is** an absolute metric
- Does not comment on
  - Granularity
  - Temporal consistency

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	

Challenge the choices that were made

- Video U-net (Ho et al. 2022b)
- V-prediction (Salimans & Ho 2022)
- Conditioning augmentation (Ho et al 2022a)
- Classifier-free guidance (Ho & Salimans 2021)
- Progressive distillation (Salimans & Ho 2022)

Challenge the choices that were made

- Video U-net (Ho et al. 2022b) — Engineered from performant image synthesis model
- V-prediction (Salimans & Ho 2022) — Empirically justified for Imagen
- Conditioning augmentation (Ho et al 2022a) — Justified from performance on image synthesis
- Classifier-free guidance (Ho & Salimans 2021) — Empirically justified for Imagen
- Progressive distillation (Salimans & Ho 2022) — Faster sampling


Imagen

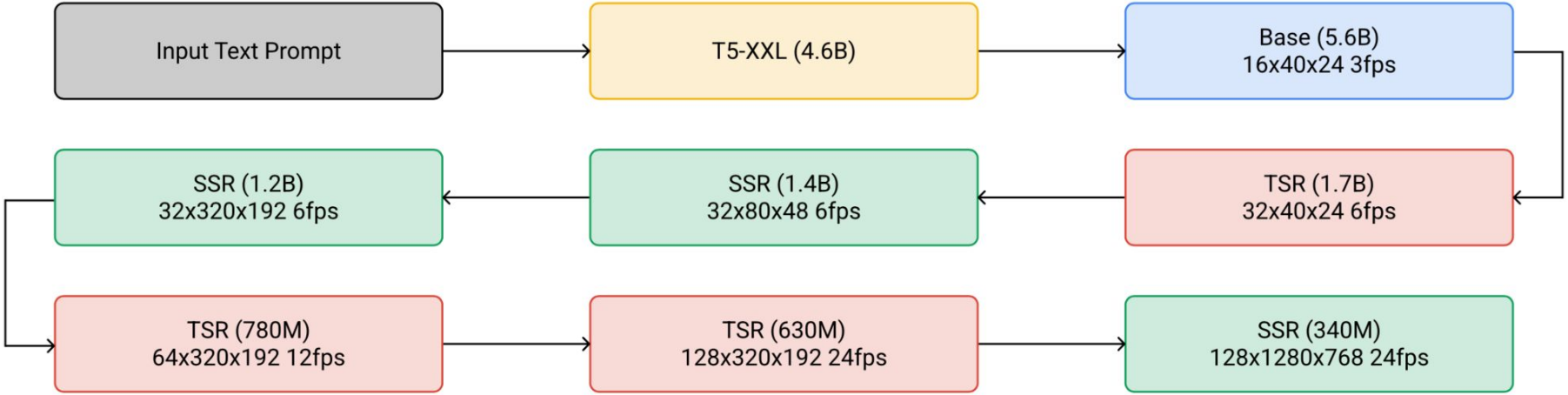

Imagen

Temporally consistent in pixel space

Inaccurate spacial super-resolution

Not consistent with objects in real space

The architecture



One more thing...

Why use diffusion?

One more thing...

Why use diffusion?

NYU

Yann Lecun's opinion

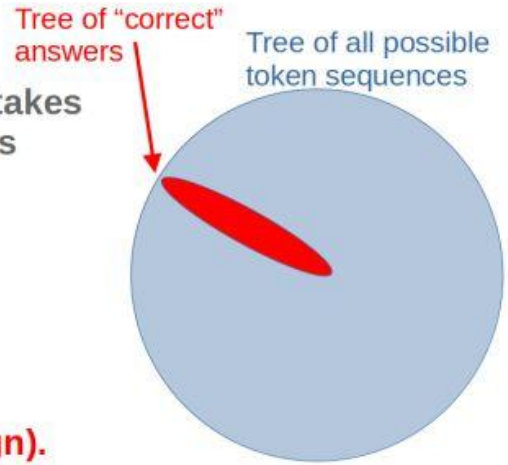
Unpopular Opinion about AR-LLMs

Y. LeCun

- Auto-Regressive LLMs are **doomed**.
- They cannot be made factual, non-toxic, etc.
- They are not controllable

- Probability e that any produced token takes us outside of the set of correct answers
- Probability that answer of length n is correct:
  - $P(\text{correct}) = (1-e)^n$

- **This diverges exponentially.**
- **It's not fixable (without a major redesign).**



Temporal attention is unlikely to fix the issue

A 3D model of a 1800s victorian house. Studio lighting.

A 3D model of a car made out of sushi. Studio lighting.

A 3D model of an elephant origami. Studio lighting.

Continuity in representation space

- Videos are 3-dimensional projection of a 4-dimensional scene
- There must be object continuity
- 3D object synthesis is a prerequisite for usable video synthesis
- It may be that Transformers can be used as a base of the pipeline