State of the Art in TTI and TTV

Armaan Khetarpaul

1 Introduction

In this, I will be evaluating the state of the art models in Text-to-Image (TTI) and Text-to-Video (TTV) generation. The state of the art is Diffusion Models, although some GAN and AutoRegressive models are also used.

2 Text-to-Image

2.1 Models in Consideration

I will be considering four models (all zero-shot):

1. Stable Diffusion v1-5 (Link)
2. DALL-E 2 (Link)(Alternate) (Unofficial Implementations)
3. Dreamlike Photoreal 2.0 (Link)
4. Imagen (Link) (Unofficial Implementation)

I have chosen these models, because they are popular and easily accessible through Hugging Face/GitHub.

2.2 Metrics

I will be using the following metrics to evaluate the models:

2.2.1 FID (automated)

FID compares the distribution of generated images with the distribution of a set of real images ("ground truth"). Lower FID scores indicate better results.

For our case, the returned images are Gaussians. The ground truth images are from the COCO (Common Objects in Context) dataset. The FID score is calculated using:

$$d_F(\mathcal{N}(\mu,\Sigma),\mathcal{N}(\mu',\Sigma'))^2 = || \mu - \mu' ||_2^2 + \mathrm{tr}(\Sigma + \Sigma' - 2(\Sigma\Sigma')^{\frac{1}{2}})$$

2.2.2 CLIP Score (automated)

CLIP Score is a reference free metric that can be used to evaluate the correlation between a generated caption for an image and the actual content of the image. It has been found to be highly correlated with human judgement. The metric is defined as:

$$CLIPScore(I, C) = \max(100 S_{\cos}(E_I, E_C), 0)$$

which corresponds to the cosine similarity between visual CLIP embedding E_I for an image I and textual CLIP embedding E_C for a caption C. The score is bound between 0 and 100 and the closer to 100 the better. $\left(S_{\cos}(x, y) = \dfrac{x \cdot y}{|| \, x \, || \, || \, y \, ||}\right)$

2.2.3 Human Score (human)

Over 1600 english prompts from the dataset PartiPrompts were used to generate images. Score given by human based on Overall alignment (1-5), Photorealism (1-5), Subject Clarity (1-5), Aesthetics (1-5), and Originality (1-5). The score is the average of the five scores. The higher the score, the better the image.

2.2.4 NSFW Score (automated)

The image is given a score from 0 to 100 based its NSFW content, Gender Bias, Skin tone bias, Toxicity and inappropriate content. These are checked by models trained on various NSFW identifying datasets. The lower the score, the better the image.

2.3 Results

The adjusted results are as follows:

Model	Lowness of FID	CLIP Score	Human Score	NSFW Score
Stable Diffusion v1-5	0.531	0.720	0.682	0.420
DALL-E 2	**0.911**	**0.983**	**0.843**	0.553
Dreamlike Photoreal 2.0	0.851	0.960	0.783	**0.300**
Imagen	0.875	0.932	0.819	0.610

The results show that DALL-E 2 is the best model in terms of FID, CLIP Score, and Human Score. Dreamlike Photoreal 2.0 is the best model in terms of NSFW Score.

3 Text-to-Video

3.1 Models in Consideration

I will be considering five models:

1. Make-A-Video (Link) (Unofficial Implementation)
2. CogVideo (Link) (Unofficial Implementation)
3. NUWA (Link) (Unofficial Implementation)
4. HiGen (Link)

5. ModelScopeT2V (Link)

I have chosen these models, because they are popular and easily accessible through Hugging Face/GitHub.

3.2 Metrics

I will be using the following metrics to evaluate the models:

3.2.1 FVD (automated)

FVD compares the generated videos with a set of real videos ("ground truth"). Lower FVD scores indicate better results. MSR-VTT dataset is used as the ground truth.

3.2.2 CLIP Score (automated)

CLIP is a reference free metric that can be used to evaluate the correlation between a generated caption for a video and the actual content of the video. It has been found to be highly correlated with human judgement. Similar to text, a video model is used here. ALternatively, average of the CLIP scores of the frames can be used.

3.2.3 FID (automated)

FID score is calculated using the same formula as in the text-to-image case, by taking the average of the FID scores of the frames. Lower FID scores indicate better results.

3.3 Results

The adjusted results are as follows:

Model	Lowness of FVD	CLIP	Lowness of FID
Make-A-Video	0.808	**0.924**	0.868
CogVideo	0.729	0.797	0.764
NUWA	0.544	0.740	0.523
HiGen	**0.865**	0.893	**0.914**
ModelScopeT2V	0.817	0.888	0.889

The results show that HiGen is the best model in terms of FVD and FID. Make-A-Video is the best model in terms of CLIP Score.