
VCLIP-Metric: A VideoCLIP-based evaluation metric for text-to-video generative tasks

Anonymous Author(s)

Affiliation

Address

email

Abstract

1 We recently witnessed the success of the text-to-video generation (T2V) models,
2 which can generate high-quality videos based on input prompts. However, eval-
3 uating the semantic correlation between the input text and the generated video
4 is challenging. In this work, we propose a VideoCLIP-based evaluation metric,
5 dubbed **VCLIP-Metric**, to capture the sequential information in the video and
6 compare its semantic information with the input text. We also collect a dataset
7 containing the generated videos from three T2V models and human evaluation of
8 the quality of the videos. The experiment on the dataset shows that our metric can
9 have a more reliable evaluation than conventional frame-based metrics.

10 1 Introduction and Background

11 The video generation task is becoming hot in the vision and language fields. Many such models are
12 devised based on approaches of Transformer [1], stable diffusion [2], or CLIP [3]. However, current
13 metrics for such text-to-video (T2V) models are limited. One of the biggest metric challenges for
14 T2V assessment is the lack of evaluating the semantic information in sequence (e.g. actions). This is
15 because current existing metrics (i) either only evaluate the texture or evaluate semantic information
16 based on frames rather than full videos; (ii) challenge to evaluate by means of capturing things like
17 actions; (iii) challenge to evaluate by means of checking whether the generated video captures all
18 contents of a sentence.

19 Most metrics work by calculating the Frechet Inception Distance score (FID) [4]. At the same time, it
20 only measures the distance between feature vectors calculated for real and generated frames (images)
21 rather than the videos. This is in contrast to the manner in which humans assess video quality, in
22 which they are always watching the whole contents of the video and then comparing it with the given
23 prompt to give a performance score eventually.

24 Although in some work, people use a new metric, Frechet Video Distance (FVD) [5], to capture the
25 temporal coherence of the content of a video in addition to the quality of each frame, it is still limited
26 as it does not evaluate the semantic information of video-text pairs and also ignores assessing whether
27 the generated videos capture all contents of the input. Therefore, it is necessary to have a metric that
28 can avoid these limitations and have a more reliable representation for the evaluation task.

29 To this end, we propose a VideoCLIP-based evaluation metric; we name it **VCLIP-Metric**, that can
30 quantitatively assess T2V generation models by explicitly comparing the semantic information in
31 sequence between input texts (prompts) and the output generated videos. VCLIP-Metric is designed
32 in light of a pre-trained VideoCLIP [6] model that is for zero-shot video and text understanding, from
33 which we extract the input prompt features and the generated video features. Further, we compare
34 the similarity of these two types of features and give a final score for the T2V models' performance.
35 Our proposed metric can also be applied to evaluate video-captioning generation tasks. In this case

36 scenario, as opposed to the input from T2V models, the input from video-captioning models would
 37 be the original videos and the post-captioning texts.

38 Due to ethical and security reasons, many T2V models do not provide open-sourced codes with
 39 pre-trained model parameters. As a result, in this work, we evaluate these three pre-trained T2V
 40 generation models: *CogVideo* [7], *Stable-Diffusion-Videos* [8], and *Real-Time Text2Video* [9].

41 2 Related Work

42 **Existing Metrics** A key challenge for the generation task is evaluating the generated contents.
 43 Generally, the metrics should measure two aspects of the generated content. The first is the texture,
 44 which examines whether the generated contents are from the same data distribution as the training
 45 samples. The second is semantics, which indicates whether the generated contents follow the input
 46 instructions. While the texture evaluation is relatively more straightforward (e.g. define similarity
 47 FID score [4], FVD scores [5], and Inception Score [5]), the semantics evaluation is more challenging
 48 because it incorporates the comparison between the semantics information in the language and vision
 49 modalities. Recently, the CLIP [3] model has shown its powerful ability to encode the text and the
 50 language input into a unified latent space that focuses on the semantics information, and achieves
 51 competitive zero-shot performance across a diverse array of image classification datasets. Thus, a
 52 straightforward evaluation metric for evaluating generated video is to average the CLIP similarity
 53 score per frame with the instruction sentence. Although there are more sophisticated CLIP-based
 54 evaluation metrics that try to more carefully examine the generation contents, such as EMScore [10],
 55 the existing metrics are based on the average similarity score between individual frames with the
 56 texts. Those approaches are inevitably challenging to capture sequential information like actions. In
 57 this work, we try to utilize the VideoCLIP [6] model to capture the video semantics to have a more
 58 accurate semantics evaluation metric.

59 **T2V models** Text-to-video (T2V) generation has been an emerging field in the last few years, but
 60 still far behind the text-to-image (T2I) generation owing to two reasons: (i) the lack of high-quality
 61 prompt-video (text-video) pairs; (ii) the complexity of modelling high-dimensional video data. The
 62 early T2V models, including SyncDraw [11], VGAN [12], and CRAFT [13], are mainly focused on
 63 simple domain video generation tasks, such as moving digits. More recently, GODIVA [14] has been
 64 developed based on 2D VQ-VAE and sparse attention for more realistic video generation. Extended
 65 from GODIVA, NUWA [15] is proposed for a variety of T2V generation tasks in a multi-task
 66 learning scheme. Furthermore, CogVideo, built by adding additional temporal attention modules on
 67 CogView [16] T2I model came up to accomplish more complicated T2V generation tasks with better
 68 performance. This year, a few state-of-the-art models emerged, Make-A-Video [17], PHENAKI [18],
 69 Imagen Video [19], Real-Time Text2Video, and Stable-Diffusion-Videos. Make-A-Video integrates
 70 the temporal U-Net with attention tensors to make predictions, and it does not require paired prompt-
 71 video data. PHENAKI tokenizer uses causal attention in time, which allows for generating varied
 72 long videos from a sequence of prompts such as a story. Imagen Video works based on a cascade
 73 of video diffusion models, which can synthesize not only high-fidelity videos but also be able to
 74 generate various artistic-style videos. Real-Time Text2Video generates videos by optimizing CLIP
 75 loss directly at the pixel level instead of optimizing through the generator model itself, leading to a
 76 fast generation speed that approaches near real-time systems. Stable-Diffusion-Videos leverages the
 77 diffusion approach, which generates videos by interpolating the latent space of stable diffusion.

78 3 Approach

79 The overall framework is described in Figure 1. Using the Video-CLIP model, we can have video
 80 and text features in the same latent space, i.e., given a video V and a text sentence T , we can get
 81 encoding f_V and f_T :

$$f_V = \text{VideoCLIP}(V), \quad (1)$$

$$f_T = \text{VideoCLIP}(T). \quad (2)$$

83 Here we try to consider the detailed information of how the input text and generated video are
 84 correlated. For example, consider a video that perfectly describes a small part of elements in the text

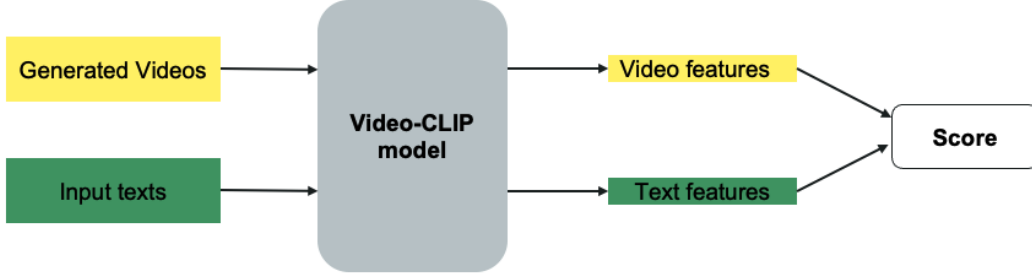


Figure 1: VCLIP-Metric Framework

but misses the majority part and a video that is like a PowerPoint to show each element per frame. We could question: is it a good generated video? We can conclude that the generated video has two desired characteristics: (a) the video describes as many elements in the text as possible, and (b) each video fragment contains as much information from sentences as possible. Inspired by the confusion matrix [20] and EMScore [10], it is useful to calculate the Recall (R) and precision (P) and get the F1 score (F) to show whether the video has these two desired characteristics. Here, we consider two sub-structures in video V and text sentence T , each 1s clips for V and semantics clusters for T (e.g. “photorealistic teddy bear”, “swimming in”, “the ocean at San Francisco”). Also, following the practice of CLIP [3], we can make the semantics cluster as a prompt (e.g. “this is a video about { }”). Assume V has m seconds and T has n semantics clusters, we have encoding features:

$$f_{V_i} = \text{VideoCLIP}(V_i) \text{ for 1s clips } i = 1, \dots, m; \quad (3)$$

$$f_{T_j} = \text{VideoCLIP}(\text{Prompt}(T_j)) \text{ for semantics clusters } j = 1, \dots, n. \quad (4)$$

Here, we cluster three type of elements in the sentence and their corresponding prompts are:

- **Object** (noun): “a video of {Object}”
- **Action** (verb): “the object in video {Action}”
- **Background** (adverbial): “the event happens {Background}”

Then we can calculate Recall (R) and precision (P) by:

$$P = \frac{1}{n} \sum_{j=1}^n \cos(f_V, f_{T_j}), \quad (5)$$

$$R = \frac{1}{m} \sum_{i=1}^m \cos(f_{V_i}, f_T). \quad (6)$$

Finally we can define our metric VCLIP-Metric score (**VCscore**) as the F1 score ($F1$):

$$\text{VCscore} = F1 = \frac{2PR}{P + R}. \quad (7)$$

Our code will be available upon publication.

4 Dataset Design

In this paper, we generated the dataset from three existing T2V models to validate our proposed metric.

V2T dataset Due to the lack of public human evaluation datasets for video generation models, we designed the dataset by ourselves with a collection of the data by ranking the prompt-video pairs from different models. More specifically, we first collected 101 prompts and inputted them into three candidate models (*CogVideo* [7], *Stable-Diffusion-Videos* [8], and *Real-Time Text2Video* [9]) to generate corresponding videos. We then ranked these generated videos from different models

112 according to the quality and extent to match the prompts. In this way, we have “ground-truth”
 113 evaluation results that can compare with the results from VCLIP-Metric, therefore validating our
 114 proposed approach. Our generated videos are 4 seconds long with 8 frames per second. The width
 115 and height are both 480 pixels.

116 Here we show an example (see Figure 2) of how to rank these models based on the video’s quality
 117 and the semantic information. The prompt for generating such videos is “**A woman wearing a red
 118 dress is smiling.**”. Each row is comprised of four frames of each generated video by different models:
 119 top by CogVideo, middle by Stable-Diffusion, and bottom by Real-Time Text2Video. Our human
 120 ranking results are that CogVideo is the best, then is Stable-Diffusion, and Real-Time Text2Video
 121 is the worst. The reason is that the video made by Cogvideo is consistent since the woman who
 122 appeared in the images remains the same, and in each frame, the woman wears a red dress and
 123 smiles. For the second video made by Stable-Diffusion, the first two images show a woman in a black
 124 dress but the last two images show a different woman in a red dress. This indicates that the model
 125 generates a discontinuous video that seems made up of a series of pictures without any context. The
 126 worst-generated video is made by the Real-Time Text2Video model, which produces a surreal video,
 127 and it is very difficult to judge if it matches the input text prompt.

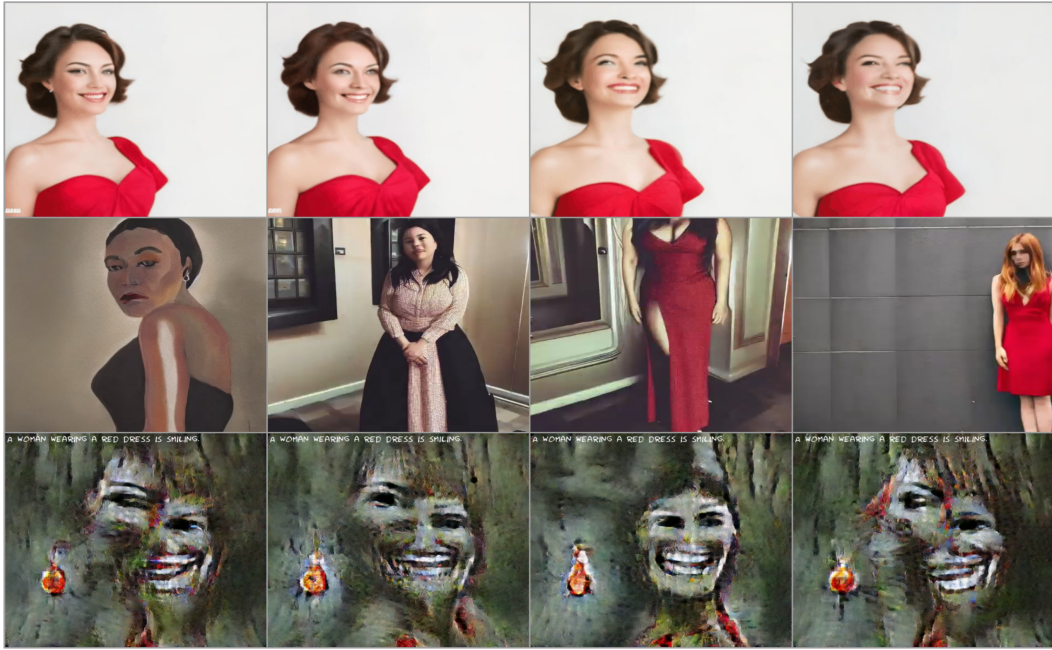


Figure 2: Rank three video generation models. Top: CogVideo; middle: Stable-Diffusion-Videos;
 bottom: Real-Time Text2Video.

128 5 Experimental Results

129 5.1 Evalutaion example

130 Here, we study an example of evaluation by our metric. The prompt is “**A man is dancing on the
 131 beach**”. We collect the generated video from the CogVideo and Stable-Diffusion model.

132 For the prompt, we cluster it into three elements as mentioned in Section 3, i.e., “**(Object)** {a man}
 133 **(Action)** {is dancing} **(Background)** {on the beach}”. And the prompts will result in:

- 134 • “a video of {a man}”
- 135 • “the object in video {is dancing}”
- 136 • “the event happens {on the beach}”

137 As shown in Figure 3, our calculated precision and recall value can help us understand the detailed
 138 correlation between input texts and generated videos.

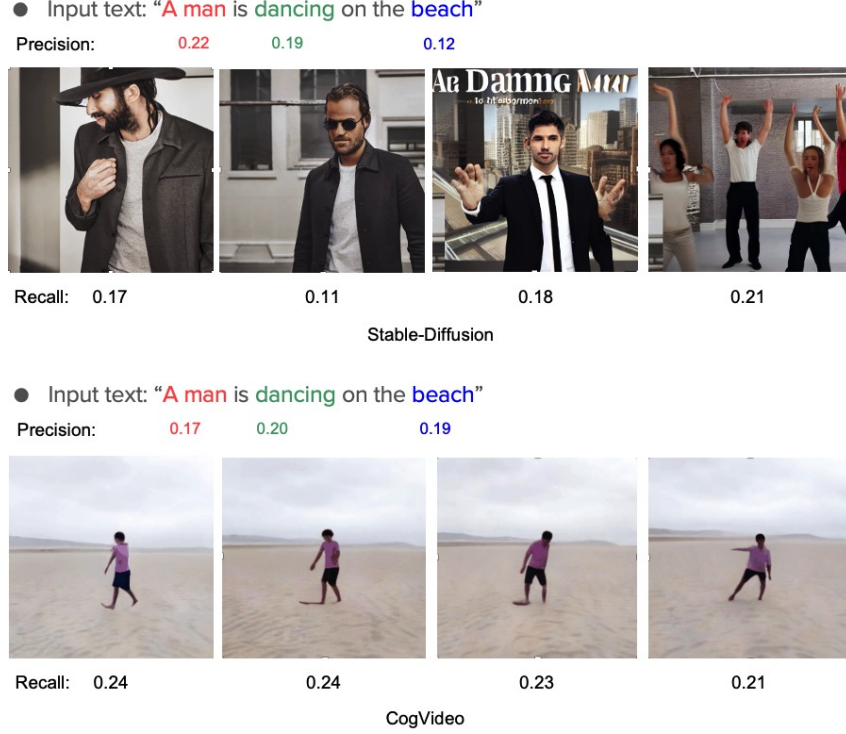


Figure 3: Example of calculated precision and recall scores for Stable-Diffusion and CogVideo model. Top: Stable-Diffusion; Bottom: CogVideo.

For the Stable-Diffusion model, the generated video is more like a set of pictures that presents each element in the sentence one by one. Therefore, its recall score is lower, especially in those 1s-clip that do not have any dancing-like actions. Also, for elements coverage (precision), it has the element of man and dancing but misses the beach. So we can see here that the precision score for the beach is much lower.

For the CogVideo model, the generated video has fluent motion, clearly showing the dynamic action of a man dancing on the beach, therefore, its recall score is higher. Moreover, because it covers all three elements in the input text, the precision score is also higher. However, the model indicates that the score for the object “a man” is lower, and indeed the quality of the man in the generated video is not high.

Overall, we can conclude that our proposed evaluation method can capture the detailed semantic correlation between input texts and generated videos, which provides a more reliable evaluation and also the interpretability of the correlation.

5.2 Correlation to human evaluation

We test the reliability of our metric by taking the human evaluation in our dataset as ground truth and calculating the Rank-order correlation between the evaluation metric scores and human evaluation scores, where the higher correlation indicates the better evaluation metric is. We compare our proposed **VCLIP-Metric** with the frame-based CLIP metric conventionally adopted in recent T2V works. The results are shown in Table 1.

Table 1: Rank-order correlation between the evaluation metric scores and human evaluation scores.

	VCLIP-Metric (ours)	CLIP frame-based metric
Rank-order Correlation	0.3465	0.1831

We find that our metric score is around 0.35, almost twice as high as the existing CLIP frame-based metric, indicating that our metric outperforms the conventional frame-based metric. It is worth noting that the CLIP frame-based metric is the most commonly used metric to assess T2V models, and almost every paper uses this metric for evaluation. The results suggest that our proposed **VCLIP-Metric** is promising and has the potential to replace the current frame-based metric for evaluating the performance of video generation models.

6 Discussion

Even though our metric is already way better than the result from the CLIP frame-based metric, it still has significant room for improvement. Here we discuss the following possible limitations and drawbacks.

6.1 CLIP struggles with abstract tasks, such as bizarre-looking videos

While CLIP usually performs well in recognizing common objects, it struggles with more abstract or systematic tasks. For example, in Figure 4, we can see that the video produced by RealTimeT2V looks not very realistic but quite horrible to a human being. In this situation, CLIP-based models are unable to recognize them, therefore, the CLIP-based metrics are no longer reliable.

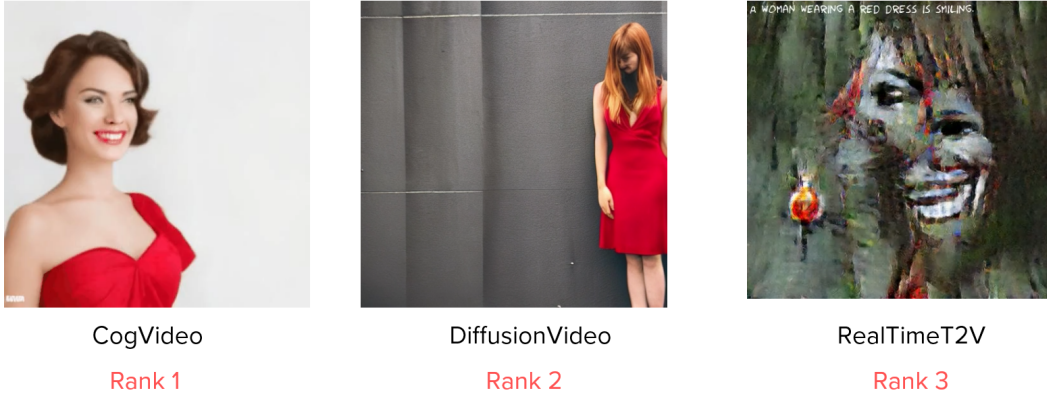


Figure 4: Prompt: A woman wearing a red dress is smiling.

6.2 T2V models we choose are not state-of-the-art

Due to the scarcity of open-source API, we did not have the video generated by some state-of-the-art models because they are non-open-sourced. This resulted in a quality drawback in our dataset.

7 Future Work

Our work opens up a few intriguing directions for further research. For instance, our metric can also assess the inverse task, i.e. video captioning. In this case scenario, we can take as input original videos and their generated captions. Then use our metric to calculate the score for comparison. Additionally, to improve the robustness of our metric, we may fine-tune the pre-trained Video-CLIP model to be suitable so that it can be suitable for various domains' videos. Furthermore, we intend to await the release of some cutting-edge T2V models, such as PHENAKI and Make-A-Video, to be public, then assess them using our metric.

References

- [1] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.
- [2] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.
- [3] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. *CoRR*, abs/2103.00020, 2021.
- [4] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.
- [5] Tim Salimans, Ian J. Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. *CoRR*, abs/1606.03498, 2016.
- [6] Hu Xu, Gargi Ghosh, Po-Yao Huang, Dmytro Okhonko, Armen Aghajanyan, Florian Metze, Luke Zettlemoyer, and Christoph Feichtenhofer. Videoclip: Contrastive pre-training for zero-shot video-text understanding. *CoRR*, abs/2109.14084, 2021.
- [7] Wenyi Hong, Ming Ding, Wendi Zheng, Xinghan Liu, and Jie Tang. Cogvideo: Large-scale pretraining for text-to-video generation via transformers, 2022.
- [8] Nathan Raw, Daniel Hug, Nirav Prajapati, Ikko Ashimine, Yiorgis Gozadinos, Aidan Holland, Jirka Borovec, and Kenneth Goodman. Create hot videos with stable diffusion by exploring the latent space and morphing between text prompts.
- [9] Peter Schaldenbrand, Zhixuan Liu, and Jean Oh. Towards real-time text2video via clip-guided, pixel-level optimization. *arXiv preprint arXiv:2210.12826*, 2022.
- [10] Yaya Shi, Xu Yang, Haiyang Xu, Chunfeng Yuan, Bing Li, Weiming Hu, and Zheng-Jun Zha. Emscore: Evaluating video captioning via coarse-grained and fine-grained embedding matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17929–17938, 2022.
- [11] Gaurav Mittal, Tanya Marwah, and Vineeth N. Balasubramanian. Sync-draw: Automatic GIF generation using deep recurrent attentive architectures. *CoRR*, abs/1611.10314, 2016.
- [12] Jiawei Chen, Janusz Konrad, and Prakash Ishwar. Vgan-based image representation learning for privacy-preserving facial expression recognition. *CoRR*, abs/1803.07100, 2018.
- [13] Tanmay Gupta, Dustin Schwenk, Ali Farhadi, Derek Hoiem, and Aniruddha Kembhavi. Imagine this! scripts to compositions to videos. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 598–613, 2018.
- [14] Chenfei Wu, Lun Huang, Qianxi Zhang, Binyang Li, Lei Ji, Fan Yang, Guillermo Sapiro, and Nan Duan. GODIVA: generating open-domain videos from natural descriptions. *CoRR*, abs/2104.14806, 2021.
- [15] Chenfei Wu, Jian Liang, Lei Ji, Fan Yang, Yuejian Fang, Daxin Jiang, and Nan Duan. Nüwa: Visual synthesis pre-training for neural visual world creation. *CoRR*, abs/2111.12417, 2021.
- [16] Ming Ding, Zhuoyi Yang, Wenyi Hong, Wendi Zheng, Chang Zhou, Da Yin, Junyang Lin, Xu Zou, Zhou Shao, Hongxia Yang, and Jie Tang. Cogview: Mastering text-to-image generation via transformers. *CoRR*, abs/2105.13290, 2021.
- [17] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiuyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, Devi Parikh, Sonal Gupta, and Yaniv Taigman. Make-a-video: Text-to-video generation without text-video data, 2022.

- 232 [18] Ruben Villegas, Mohammad Babaeizadeh, Pieter-Jan Kindermans, Hernan Moraldo, Han Zhang,
233 Mohammad Taghi Saffar, Santiago Castro, Julius Kunze, and Dumitru Erhan. Phenaki: Variable
234 length video generation from open domain textual description, 2022.
- 235 [19] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko,
236 Diederik P. Kingma, Ben Poole, Mohammad Norouzi, David J. Fleet, and Tim Salimans.
237 Imagen video: High definition video generation with diffusion models, 2022.
- 238 [20] David MW Powers. Evaluation: from precision, recall and f-measure to roc, informedness,
239 markedness and correlation. *arXiv preprint arXiv:2010.16061*, 2020.