A Dataset of Text Prompts, Videos and Video Quality Metrics from Generative Text-to-Video AI Models

Iya Chivileva[1], Philip Lynch[1], Tomás E. Ward[2] and Alan F. Smeaton[2] (*)

[1] School of Computing, Dublin City University, Glasnevin, Dublin 9, Ireland
[2] Insight Centre for Data Analytics, Dublin City University, Glasnevin, Dublin 9, Ireland

Alan.smeaton@dcu.ie @asmeaton

Keywords
Generative AI; Video annotation; Video naturalness; Video perception; Video alignment;

Abstract

Evaluating the quality of videos which have been automatically generated from text-to-video (T2V) models is important if the models are to produce plausible outputs that convince a viewer of their authenticity. This paper presents a dataset of 201 text prompts used to automatically generate 1,005 videos using 5 very recent T2V models namely Tune-a-Video, VideoFusion, Text-To-Video Synthesis, Text2Video-Zero and Aphantasia. The prompts are divided into short, medium and longer lengths. We also include the results of some commonly used metrics used to automatically evaluate the quality of those generated videos. These include each video's naturalness, the text similarity between the original prompt and an automatically generated text caption for the video, and the inception score which measures how realistic is each generated video.

Each of the 1,005 generated videos was manually rated by 24 different annotators for alignment between the videos and their original prompts, as well as for the perception and overall quality of the video. The data also includes the Mean Opinion Scores (MOS) for alignment between the generated videos and the original prompts.

The dataset of T2V prompts, videos and assessments can be reused by those building or refining text-to-video generation models to compare the accuracy, quality and naturalness of their new models against existing ones.

SPECIFICATIONS TABLE

Subject	Computer Vision and Pattern Recognition
Specific subject area	Dataset for the evaluation of text-to-video generation models which automatically generate videos from text prompts.
Data format	Raw videos as .mp4 and .gif; text prompts as .txt; video metric scores and annotations as .csv.
Type of data	Text prompts in .txt file, videos in .mp4 and in .gif formats, .CSV file of scores for video naturalness, automatic measures of similarity between the prompts and videos and summary scores from human annotations of video quality and alignment between prompts and videos.

Data collection	Text prompts were selected by combining content generated by ChatGPT with manual curation. Video data was created by using the text prompts as inputs to 5 popular text-to-video generation models. The resulting videos were analysed using metrics for their quality and for the naturalness as well as their similarity to the prompts used to generate them. Human annotators then rated the videos for quality and for closeness to their original prompts, using an online annotation tool.
Data source location	Annotators who performed manual video annotations online were based in the area around Dublin, Ireland. All other data was generated from computer processing.
Data accessibility	Repository name: Figshare Data identification number: 10.6084/m9.figshare.24078045.v3 Direct URL to data: https://doi.org/10.6084/m9.figshare.24078045.v3 GitHub URL for code for implementing video naturalness calculation: https://github.com/Chiviya01/Evaluating-Text-to-Video-Models

VALUE OF THE DATA

- This first-of-its-kind dataset can be reused by researchers working on text-to-video generation models who wish to evaluate the accuracy and naturalness of their own generated videos. Researchers can re-use the prompts from this data to generate their own videos and compute results from automatic metrics for naturalness, quality and alignment with the text prompt as well as re-using the code provided to calculate video naturalness, and then compare those results with the metric results from videos in this dataset.
- The dataset may be used to train text-to-video and/or video-to-text deep learning models.
- The text prompts vary in length with 87 of them being short (4 to 8 words), 43 are of average length (9 to 13 words) and 71 are longer than 13 words and this allows researchers to explore the relationship between prompt length and the quality of a generated video.
- The dataset gives insights into the comparative performances of 5 popular text-to-video models, namely Tune-a-Video, VideoFusion, Text-To-Video Synthesis, Text2Video-Zero and Aphantasia.
- The dataset is freely available for public download.

BACKGROUND

In the field of AI-generated images, recent work [1] has compared the quality of images generated from a collection of text-to-image (T2I) models. This used automatically-computed metrics for image quality as well as human evaluations of perception and alignment of the image to the text prompt used to generate it. That work also included the release of a database of images and the prompts and model parameters used to generate them as well as human evaluations and outputs from automatic metrics.

The creation and release of the dataset described here follows a similar sequence to [1] except we address evaluating the quality of text-to-video (T2V) instead of text-to-image generative models. The dataset includes the videos generated from each of 5 popular T2V models using the same text prompts, as well as metric values for video quality, naturalness, perception and alignment. It also includes human annotations of those generated videos which also measure video quality and alignment between the text prompt and the generated video.
The creation of the dataset allows researchers to directly compare the performance of their own text-to-video (T2V) models against others from the literature, using a common dataset, common automatically-computed metrics and human annotations.

DATA DESCRIPTION

The dataset comprises 201 text prompts which were used in 5 T2V models to generate 1,005 T2V model videos. We carefully selected the 201 prompts by combining content generated by ChatGPT with manual curation. The compilation covers a broad range of topics including influential figures, notable places, and cultural events like Easter and the Brazilian Carnival. 87 of the prompts are short (4 to 8 words), 43 are of average length (9 to 13 words) and 71 are longer than 13 words. The collection of prompts offers a diverse range of videos, spanning from practical scenarios to creative concepts. The videos encompass a variety of actions, relationships, and visual styles. Sample frames from the collection of generated videos are shown in Figure 1.

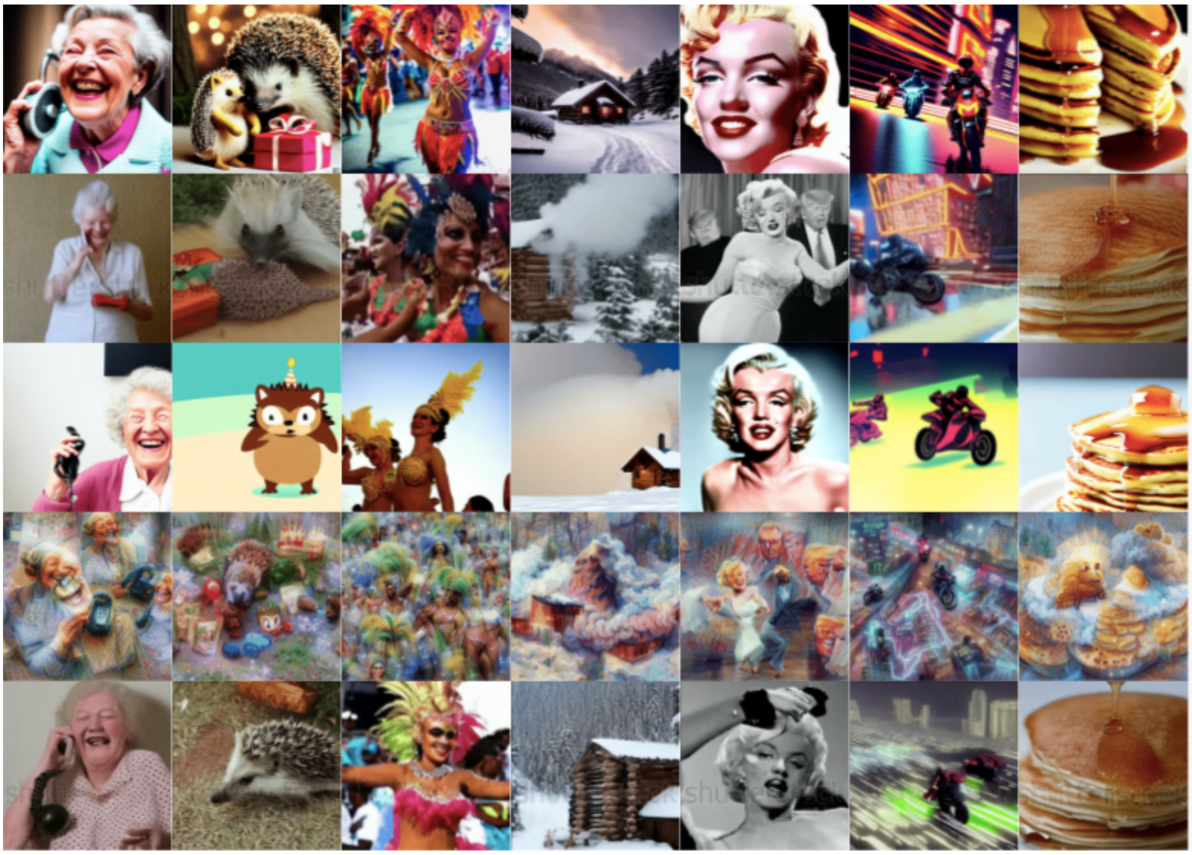

Figure 1 Samples from our generated videos -- rows show frames generated by Text2Video-Zero, Text-to-Video Synthesis, Tune-a-Video, Aphantasia and Video Fusion respectively while the columns are frames from the same text prompts.

The text prompts are in the file "text_prompts_for_video_generation.txt" with one prompt per line, each ended with a full stop.

The videos generated from each of 5 T2V models are in MP4 format except for those generated from Tune-a-Video which are animated GIFs. The 1,005 videos are grouped into 5 Zipped folders, one for each model, and the expanded Zipfiles are named Aphantasia, Tune-a-video, Video_Fusion, ModelZero (for Text2Video-Zero) and T2VSynthesis respectively. The final file in the dataset is "T2V Numerical Data.csv" which contains a row entry for each of the 1,005 videos with a column for each of the "unique_video_name", the "original prompts", the "model" used to generate the video (one of Aphantasia, ModelZero, T2VSynthesis, Video_Fusion or Tune-a-Video), the "prompt length" (one of long, short or average), the "naturalness_score" for the video (in the range 0 to 1.0), a "text_similarity" score (in the range 0 to 1.0) between the originating prompt and an automatically-determined caption for the generated video which is a similarity algorithm described later, the "BLIP_SIM" similarity (in the range 0 to 1.0) between the originating prompt and the automatically-determined caption, the "inception score" and the mean opinion score (MOS) from the anonymized 24 manual annotations of the video for both alignment between the originating prompt and the generated caption ("MOS Alignment"), and for the perceptual quality of the generated video ("MOS Perception"). The final column contains the average of the two MOS values.

EXPERIMENTAL DESIGN, MATERIALS AND METHODS

In 2022 the first open-source T2V model called **Tune-a-Video** was released by Wu *et al.* [2] introducing a mechanism that uses the Stable Diffusion model [3] for video generation. This model is built on state-of-the-art T2I diffusion models and involves a tailored spatio-temporal attention mechanism and an efficient one-shot tuning strategy. It served as an inspiration for the rapid development of other open-source models including **VideoFusion** [4] which, in 2023, uses a decomposed diffusion process to resolve per-frame noise as a base noise that is shared among all frames leading to smoother video output. **Text-to-Video Synthesis**, also based on the work described in [4] in 2023 is also a multi-stage text-to-video generation diffusion model which consists of text feature extraction, a text feature-to-video latent space diffusion model, and video latent space to video visual space. **Text2Video-Zero** [5] in 2023 takes a low-cost zero-shot text-to-video generation approach, leveraging the power of Stable Diffusion and tailoring it for video. Finally **Aphantasia** [6] also from 2023, is a collection of text-to-image tools, evolved from the artwork of the same name which also generates video format outputs.

Having generated a total of 1,005 videos from the set of 201 prompts we then calculated the values of a number of automatic and manual metrics to enhance the usefulness of the dataset. One of the most commonly used automatic metrics for evaluating video quality is Inception Score (IS) [7] which was developed as an alternative to human evaluation and aims to measure both image quality and diversity. It relies on the ``inception network" [8] to generate a class probability distribution for images and IS scores for teach video are included in this dataset.

Image naturalness refers to how realistic and free of distortions or artefacts an image appears. Naturalness is related to quality, which encompasses aspects such as sharpness, contrast, and colour accuracy, but naturalness specifically focuses on the realism of an image.
We developed and applied a classifier for video naturalness for which we collected several statistical measures from each video, including:

- Texture score measures the degree of uniformity in a video frame's texture since natural images, such as landscapes or animal fur, tend to have more complex textures than synthetic images. After converting to grayscale and applying a Gaussian blur to reduce noise we apply Sobel edge detection in the x and y directions and calculate the magnitude of the gradient. The variance of this magnitude is the texture score.
- The sharpness score measures the amount of high-frequency content in a video frame, indicative of the image's level of detail. It is calculated by applying a sharpening filter to the image and then taking the RMS difference between the original and the filtered image.
- The colour distribution score is a measure of the uniformity of colour in a frame, exploiting the characteristic of a uniform or artificial colour distribution in a non-natural image. It is calculated by applying K-means clustering with K=2 to the A and B channels of the frame's LAB representation. This score is the proportion of pixels in the cluster with the lowest A channel value.
- The spectral score measures the extent to which a frame differs from the natural image statistics in the Fourier domain. The function calculates the mean and standard

deviation of each colour channel and computes the spectral score as the sum of standard deviations divided by the sum of means.

- The entropy score uses the Shannon entropy formula [9] which measures the level of randomness or disorder in pixel values. Natural images tend to have a higher degree of order and lower entropy than non-natural ones.
- The contrast score measures differences between the lightest and darkest parts of a video frame by dividing the standard deviation of pixel intensities by the mean intensity.
- Oriented FAST and Rotated BRIEF (ORB) is a feature detection algorithm [10] to compute statistics about the key points in a frame including the mean and standard deviations of the distances between key points and of the lengths of the descriptors associated with those key points.
- The number and sizes of blobs is detected using the Laplacian of Gaussian (LoG) method [11]. Blobs are regions in a video frame with a relatively uniform intensity that stand out compared to the surrounding area.
- The Naturalness Image Quality Evaluator (NIQE) is a no-reference image quality assessment metric [12] based on the observation that natural images tend to exhibit a unit-normal Gaussian characteristic in their luminance values. NIQE uses a set of natural scene statistics (NSS) that captures the statistical regularities present in natural scenes that are not present in unnatural or distorted images.

To enable processing, a YUV444 video frame is reshaped from planar to interleaved format, which represents colour information in terms of brightness (Y) and colour (U and V), with 8 bits allocated to each channel. NIQE scores were calculated for the grayscale frames and for the Y, U and V channels in the YUV444 video frames separately as this provides a better visual representation of the image [13].

To train the classifier for video naturalness we also calculated a Modified Inception Score (MIS) for each video which operates on a similar principle to Inception Score mentioned earlier by calculating the mean probability distributions of all frames in a generated video. We modified the IS metric to return a larger value if the mean probability distribution in a video has low entropy. Essentially, if the Inception model assigns a greater probability to one particular class throughout the frames in a video, MIS will produce a larger value. We achieved this by setting the marginal distribution to the uniform distribution.

We collected all video feature data described above from 187 videos comprising 92 natural and 95 non-natural scenes. We approached the naturalness classifier task as a binary classification problem and manually assigned each video a label indicating natural or not. We trained three classifiers, AdaBoost, a Bagging classifier with a DecisionTree base and XGBoost. To optimise the performance of each classifier, we employed GridSearch. We evaluated the classifiers' performance using F1 on training, validation, and test sets. The XGBoost classifier performed the best on unseen data and was used to calculate the values of naturalness score for each video in the dataset.

To measure alignment between the original text prompt and the generated video we measured the semantic similarity between captions for the generated videos and the original text prompts. The process involves generating captions for each video frame using the BLIP-2 [14] image caption generator. In our approach to measuring alignment for generated videos we combine BERT and Cosine similarities.

BERT (Bidirectional Encoder Representations from Transformers) models [15] measure the similarity between two pieces of text. BERT is designed to capture more nuanced and complex semantic relationships between sentences or between captions and prompts in this case, whereas Cosine similarities only consider surface-level similarity based on word overlap. Our similarity metric penalises the BERT similarity score with the Cosine similarity score ensuring that the combined similarity shown in Equation 1 reflects both surface-level and deeper semantic similarities between captions and prompts. After multiple experiments we determined the optimal ratio between BERT and Cosine similarities to be 0.75:0.25.

$$Text\ Similarity = \begin{cases} 0.25\ (Cos\ sim) + 0.75\ (BERT\ sim), & if\ Cos\ sim \neq 0 \\ 0.5\ (BERT\ Sim) & otherwise \end{cases}$$

Eq.1

Given that some frames in generated videos may exhibit significant distortions or omissions or not contain recognisable objects such as in Figure 2 where two frames in a generated video do not include a dog, we calculate the weighted textual similarity for a generated video of n frames as $\frac{1}{n}\sum_{i=1}^{n} w_i.sim_i$. The weights are assigned based on the frequency of each caption in the overall list of generated captions and these are the values that appear in the "text_similarity" column in the dataset.

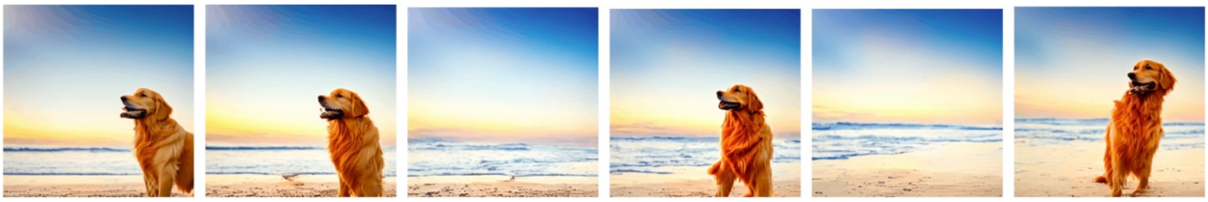

Figure 2: Selected frames from a generated video with the prompt "A golden retriever eating ice cream on a beautiful tropical beach at sunset". Note that 2 of the frames are missing the dog.

To obtain human quality evaluation scores for the generated videos we recruited volunteers to rate videos remotely and in their own time with each person given 10 days to complete the task across up to 10 interactive sessions. The annotators rated each video on a scale of 1 (low) to 10 (high) for two aspects, alignment and perception. They were asked to rate each video on a scale of 1 to 10 for each of two categories and were given the following definitions of alignment and of perception:

- "Alignment Score reflects the compatibility between the generated video and the text of the original prompt so consider all elements in the text as crucial."
- "Perception Score rates the perceptual quality of the video taking into account issues such as visual clarity: How clear and sharp are the visuals in the video clip? Are there any visual artifacts, blurriness, or pixelation? How accurately do the colours represent the real-world scene or intended visual style as some of these are real world, others cartoon, others impressionist, etc.? Are there any colour distortions or inaccuracies? What is your overall impression of the quality of the video produced?"

24 (16 male, 8 female) adult annotators, mostly graduate students, completed two ratings of each video giving 1,005 videos x 24 annotators x 2 ratings = 48,240 quality ratings. Annotators were rewarded with a gift token when they completed annotating the videos.
In assessing the quality of still images the ``de facto" metric is mean opinion score (MOS) [16] which is the mean of the opinions and ratings of human evaluators gathered according to

some numeric or qualitative scale such as we use here. In [17] the authors proposed that the standard deviation of opinion scores reflects the subjective diversity while more recently [18] proposed that as well as the mean of the opinion scores, researchers should assess quality in terms of the distribution of opinion scores, not just the standard deviation.

Figure 3 shows the distribution of adjusted MOS scores for alignment and perception for all 1,005 videos while Figure 4 shows the distributions on a per-model basis. Table 1 shows the mean and standard deviations of MOS scores for alignment and perception for videos generated by each of the 5 models.

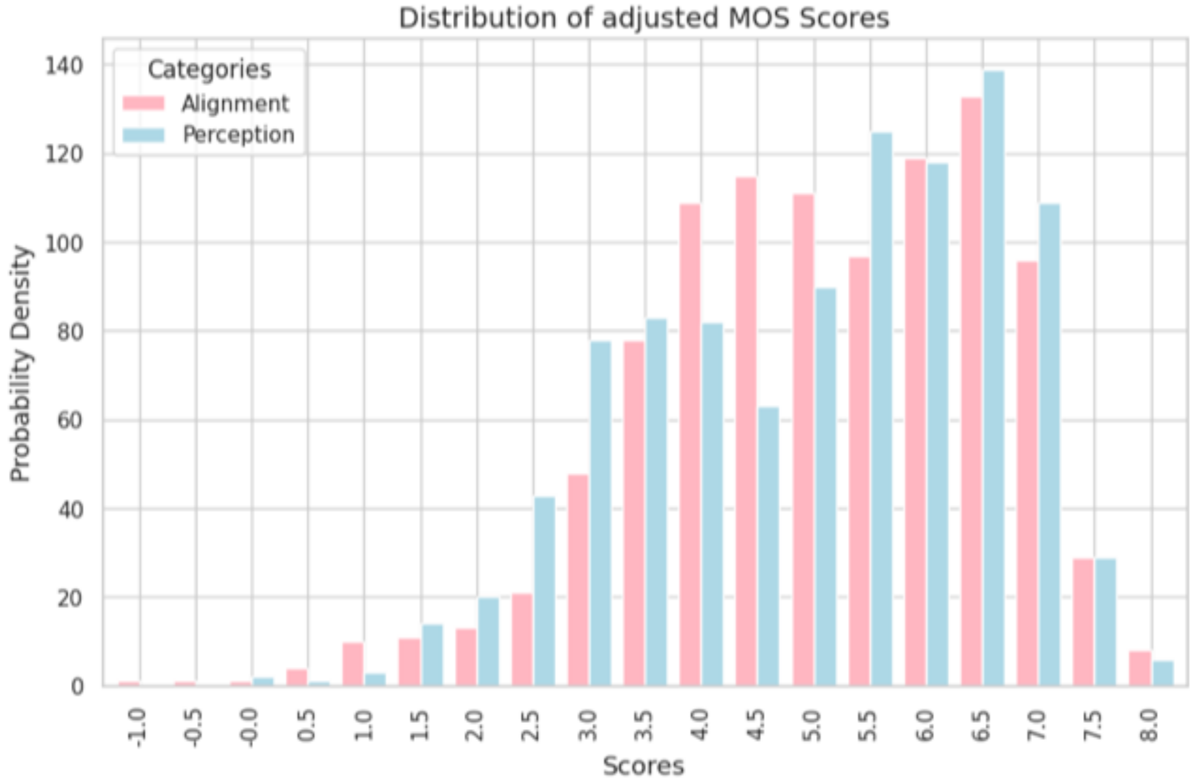

Figure 3: Distribution of adjusted MOS Scores.

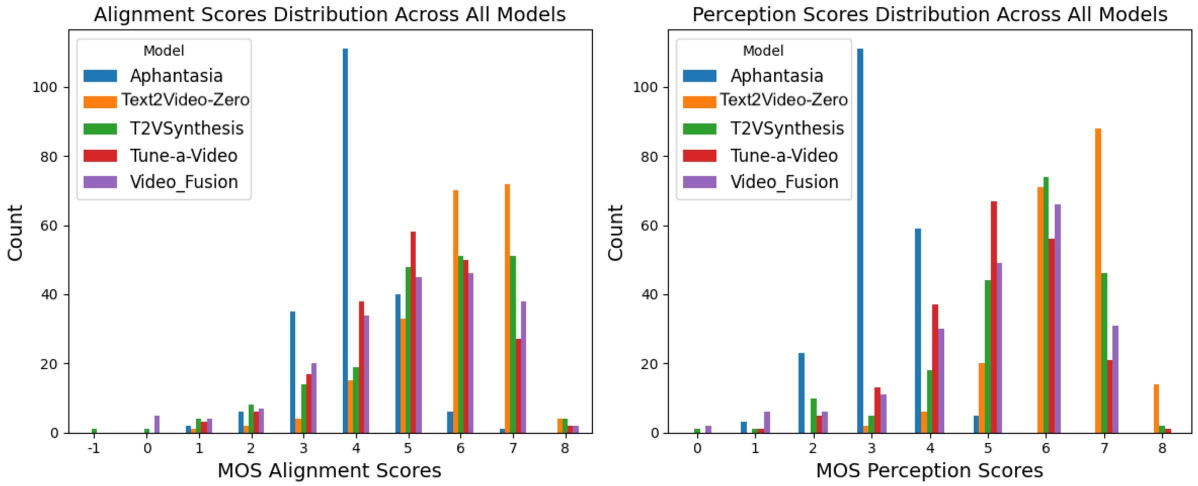

Figure 4: MOS score distributions across different models.

Table 1: Human Evaluation MOS Scores for five Text-to-Video Models.

Model	Alignment		Perception	
	Mean	**Std.Dev.**	**Mean**	**Std.Dev.**
Aphantasia	4.016	0.841	3.221	0.692
Text2Video-Zero	5.985	1.139	6.393	0.886
T2VSynthesis	5.333	1.622	5.485	1.366
Tune-a-Video	5.053	1.340	5.070	1.196
Video Fusion	4.995	1.686	5.139	1.507

LIMITATIONS

None.

ETHICS STATEMENT

This research was carried out in accordance with the Declaration of Helsinki, and includes ethical approval from the School of Computing Research Ethics Committee 2023-01-20. Informed consent was obtained from human participants who provided video annotations.

CRediT AUTHOR STATEMENT

Iya Chivileva: Conceptualization, Methodology, Software, Validation, Formal analysis, Writing - Original Draft, Visualization. **Philip Lynch:** Conceptualization, Methodology, Software, Validation, Formal analysis, Writing - Original Draft, Visualization. **Tomás E. Ward:** Methodology, Validation, Data Curation, Writing - Review & Editing, Supervision. **Alan F. Smeaton:** Conceptualization, Methodology, Validation, Data Curation, Writing - Review & Editing, Supervision.

ACKNOWLEDGEMENTS

This research was funded by Science Foundation Ireland (SFI) under Grant Number: SFI/12/RC/2289_P2, co-funded by the European Regional Development Fund.

DECLARATION OF COMPETING INTERESTS

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

REFERENCES

1. Li, C., Zhang, Z., Wu, H., Sun, W., Min, X., Liu, X., ... & Lin, W. (2023). AGIQA-3K: An Open Database for AI-Generated Image Quality Assessment. arXiv preprint arXiv:2306.04717. https://doi.org/10.48550/arXiv.2306.04717

2. Wu, J. Z., Ge, Y., Wang, X., Lei, S. W., Gu, Y., Shi, Y., ... & Shou, M. Z. (2023). Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 7623-7633).

3. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2022). High-resolution image synthesis with latent diffusion models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 10684-10695). https://doi.org/10.1109/CVPR52688.2022.01042

4. Luo, Z., Chen, D., Zhang, Y., Huang, Y., Wang, L., Shen, Y., ... & Tan, T. (2023). VideoFusion: Decomposed Diffusion Models for High-Quality Video Generation. arXiv e-prints, arXiv-2303. https://doi.org/10.48550/arXiv.2303.08320

5. Khachatryan, L., Movsisyan, A., Tadevosyan, V., Henschel, R., Wang, Z., Navasardyan, S., & Shi, H. (2023). Text2video-zero: Text-to-image diffusion models are zero-shot video generators. arXiv preprint arXiv:2303.13439. https://doi.org/10.48550/arXiv.2303.13439

6. Vadim Epstein. Aphantasia text to video model. https://github.com/eps696/ aphantasia. Last Accessed: 29 August 2023.

7. Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., & Chen, X. (2016). Improved techniques for training GANs. *Advances in neural information processing systems*, *29*. https://doi.org/10.5555/3157096.3157346

8. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2016). Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2818-2826). https://doi.org/10.1109/CVPR.2016.308

9. Cover, T. M., & Thomas, J. A. (2006). Elements of information theory second edition solutions to problems. *Internet Access*, 19-20.

10. Rublee, E., Rabaud, V., Konolige, K., & Bradski, G. (2011, November). ORB: An efficient alternative to SIFT or SURF. In *2011 International conference on computer vision* (pp. 2564-2571). IEEE. https://doi.org/10.1109/ICCV.2011.6126544

11. Lindeberg, T. (1998). Feature detection with automatic scale selection. *International Journal of Computer Vision*, *30*, 79-116. https://doi.org/10.1023/A:1008045108935

12. Mittal, A., Moorthy, A. K., & Bovik, A. C. (2012). No-reference image quality assessment in the spatial domain. *IEEE Transactions on Image Processing*, *21*(12), 4695-4708. https://doi.org/ 10.1109/TIP.2012.2214050

13. Podpora, M., Korbas, G. P., & Kawala-Janik, A. (2014, October). YUV vs RGB-Choosing a Color Space for Human-Machine Interaction. In *FedCSIS (Position Papers)* (pp. 29-34). https://doi.org/ 10.15439/978-83-60810-57-6

14. Li, J., Li, D., Savarese, S., & Hoi, S. (2023). Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*. https://doi.org/10.48550/arXiv.2301.12597

15. Reimers, N., & Gurevych, I. (2019, November). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (pp. 3982-3992). https://doi.org/10.18653/v1/D19-1410

16. Streijl, R. C., Winkler, S., & Hands, D. S. (2016). Mean opinion score (MOS) revisited: methods and applications, limitations and alternatives. *Multimedia Systems*, *22*(2), 213-227. https://doi.org/10.1007/s00530-014-0446-1

17. Hoßfeld, T., Schatz, R., & Egger, S. (2011, September). SOS: The MOS is not enough!. In *2011 third international workshop on quality of multimedia experience* (pp. 131-136). IEEE. https://doi.org/ 10.1109/QoMEX.2011.6065690

18. Gao, Y., Min, X., Zhu, Y., Li, J., Zhang, X. P., & Zhai, G. (2022, October). Image quality assessment: From mean opinion score to opinion score distribution. In *Proceedings of the 30th ACM International Conference on Multimedia* (pp. 997-1005). https://doi.org/10.1145/3503161.3547872