Automatic Text-To-Video Model Evaluation

Iya Chivileva and Philip Lynch

Dublin City University, Collins Ave Ext, Whitehall, Dublin 9, Ireland
{iya.chivileva2, philip.lynch35}@mail.dcu.ie

Abstract. Our research aims to advance the field of Text-to-Video (T2V) generation through the development of a novel evaluation metric. The study highlights the constraints of the current evaluation metrics used in the literature, identifying the need for a new evaluation metric that addresses these limitations. Our main contribution is a proposed evaluation metric which involves addressing two critical challenges: image naturalness and semantic matching. Our research demonstrates that the novel metric outperforms a commonly used metric, indicating that it is a reliable and valid evaluation tool. The proposed evaluation technique enhances the existing evaluation frameworks and provides a more reliable measure of the quality of generated videos, thus improving the quality and relevance of T2V technology. The code and examples used in the study are available for public access[1].

Keywords: Text-to-Video · Video Processing · Evaluation

1 Introduction

This project is structured into three main parts, with a focus on advancing the field of T2V generation through the development of a novel evaluation metric. Due to the rapid advancements in T2V algorithms, a majority of the research is concentrated on improving the SOTA while producing a reliable metric is more of an afterthought. In literature, it is common for a model to be evaluated on 3 or 4 different metrics and in some cases human evaluation. Our goal was to create a metric which can replace the expensive and slow process of human evaluation and reduce the number of metrics needed to produce a reliable evaluation of a T2V model.
In the first part of the paper, we conducted a thorough examination of the state-of-the-art (SOTA) T2V models. In section 2.1, we provide an extensive review of both open-source and close-source models in the current literature. Given the rapid pace of development in this field, new models are continually emerging, making it essential to stay abreast of the latest advancements. In the second part of our research, we focused on the constraints of the current evaluation metrics used in the literature. Specifically, in section 2.2, we highlight the major pitfalls of each metric and identify the need for a new evaluation metric that addresses these limitations. In the third and most critical part of the study, we present our proposed evaluation metric, which builds on the limitations of the

[1] GitLab

current evaluation metrics. Our proposed approach tackles three significant challenges that existing metrics have failed to adequately address. The first challenge is assessing the authenticity and realism of a video, which we refer to as video naturalness (discussed in Section 2.4). We developed a reliable classifier to detect video naturalness, as presented in Section 3.1. The second challenge is evaluating the degree to which the generated video aligns with the given input prompt, also known as text similarity (discussed in Section 3.2). The third challenge is semantic matching between the original text-prompt and the content inside the generated video. We propose a novel semantic matching evaluation metric in Section 3.3. Finally, in Section 4, we analyze and compare the results of our novel metric with a commonly used T2V metric to produce a highly correlated score with human interpretation. Our findings suggest that our novel metric outperforms the commonly used metric, indicating that it is a reliable and valid evaluation tool.
The code and examples used in this study are available at gitlab/2023-ca4021-PLynch-IChivileva, including instructions on how to run models from video generation and evaluation.

2 Related Works

2.1 Text-to-Video Models

We began our research by exploring the SOTA T2V models, through our research we found that a majority of the models are closed-sourced and initially, the biggest problem that we faced was finding open-source T2V models but our biggest breakthrough came when Wu et al. released their model Tune-a-Video[25], they introduced a mechanism that utilises the Stable Diffusion model [16] for video generation. This model served as an inspiration for developing open-source models such as VideoCrafter[2], Text2Video-Zero [5], VideoFusion [9] and Text-to-Video Synthesis[3]. These models were utilized in our evaluation, as they constitute the bulk of the videos upon which we assessed our metrics.

2.2 Evaluation Metrics

This section provides an overview of the most commonly used video evaluation metrics in the current literature. Each metric is introduced with the original paper in which it was proposed, followed by a brief explanation of its objective and how it functions.
Inception Score(IS) [18]was developed as an automated alternative to human evaluation, with the goal of reducing costs, increasing speed, and eliminating the biases associated with human evaluation. The metric aims to measure both image quality and diversity. It relies on the "inception network" [21] to generate the class probability distribution for images. Higher-quality images should have

[2] VideoCrafter GitHub
[3] Text-to-Video Synthesis ModelCard

a low entropy probability $p(y|x)$, while diversity is measured with the marginal distribution of all the images, which should have high entropy as illustrated in Figure 1.

Fig. 1: IS Ideal Image Probability Distribution [10]

Fréchet Video Distance(FVD) [22] measures the distance between the feature activations of real and generated videos in the feature space of a pre-trained video classifier. Similar to the Fréchet Inception Distance (FID) [3] which was developed for images, FVD measures the distance between the distributions of real and generated videos. A lower FVD score indicates better quality video generation, as this indicates that both real and generated videos have similar distributions for their feature activations.
CLIPSim [24] uses the CLIP [15] model to evaluate the semantic matching between the initial text prompt and the generated video. CLIP is a contrastive learning-based model that creates a joint embedding space for images and text, allowing the model to understand the relationships between them. CLIPSim extends CLIP to evaluate videos by finding the CLIP score of each frame in a video and returning the average score of all the frames in a video.
Although this list is not exhaustive, the metrics described are among the most widely used and accepted metrics for evaluating T2V models.

2.3 Potential Drawbacks

In this section, we address some concerns related to the commonly used evaluation metrics. Despite their usefulness, it is important to acknowledge that no metric is perfect and each has its limitations. We highlight some major concerns that we have encountered with the metrics discussed earlier.
Firstly, **IS** has been criticized for its tendency to overfit on models trained using ImageNet and in some cases its inability to distinguish between poor and high-quality images, as demonstrated in Figure 2a where IS assigned an almost perfect score to these examples. While these concerns were raised by Barratt et al. (2018) [1], many researchers still rely on IS as a metric.
Secondly, **FVD** requires real videos to generate a score, which poses a challenge in fairly comparing different T2V models that are trained on different datasets. This limitation hinders the broader applicability of this metric.
Thirdly, the underlying model used in CLIPSim, CLIP, has been criticized by the authors of BLIP (2022) [6] for its reliance on noisy web image-text pairs. They argue that a smaller, filtered dataset should be used instead. Furthermore, the ability of Image-to-Text models such as BLIP and CLIP to generate

semantically similar captions for images/frames that do not visually appear to match is another concern, as illustrated in Figure 2b, the video was generated using the Aphantasia T2V model[4] with the text-prompt "A blue unicorn flying over a mystical land" and by using CLIPSim we generated a similarity score of over 70%.
In summary, while the metrics discussed in this section are widely used and accepted, they are not without their limitations. These are some of our own motivations for our research into possible solutions to these drawbacks but we also urge researchers to continue devoting significant effort to the evaluation and validation of new techniques and methods, in order to ensure the effectiveness of the metrics used.

(a) Poor Quality Images [1]

(b) CLIPSim Error

Fig. 2: Examples of Metric Limitations

2.4 Image Naturalness

Image naturalness refers to how realistic and free of distortions or artefacts an image appears. It is closely related to image quality, which encompasses aspects such as sharpness, contrast, and colour accuracy, but image naturalness specifically focuses on the realism of an image. Non-natural images are those that lack recognizable and interpretable real-world objects or scenes. These images may include computer-generated graphics, abstract art, or heavily manipulated photographs, and they often serve artistic or functional purposes, but they do not necessarily reflect the properties of natural images that are easily interpreted by human vision. The following are metrics that are currently available for measuring the naturalness of images:
VIF (Visual Information Fidelity) is a full-reference image quality assessment (IQA) metric [19]. The VIF metric takes into account the visual sensitivity of human observers to structural distortions in the image, which is based on the properties of the Human Visual System (HVS). The VIF metric compares the statistics of gradient magnitudes of the distorted and reference images in a local spatial neighbourhood, which is specified by the user. The similarity

[4] Aphantasia Model

between the two images is measured by comparing the normalized covariance between the gradients of the distorted image and the reference image to the product of the normalized variances of the gradients in each image. The VIF metric is specifically designed to assess the quality of distorted natural images, and it requires a reference image that closely represents the original undistorted image [23].

SSIM (Structural Similarity Index) is a full-reference image quality assessment metric that measures the structural similarity between the distorted image and the original reference image. It takes into account the sensitivity of human visual perception to changes in structural information, luminance, and contrast of the image [19]. The SSIM metric's accuracy can be affected by the quality of the reference image, according to a study [23] on its performance in the presence of various distortions and noise.

The Naturalness Image Quality Evaluator (NIQE) is a no-reference image quality assessment metric [12]. NIQE is based on the observation that natural images tend to exhibit a unit-normal Gaussian characteristic in their luminance values. Therefore, NIQE utilizes a set of natural scene statistics (NSS) that capture the statistical regularities present in natural scenes that are not present in unnatural or distorted images, including those related to image contrast, luminance, colour, and texture. The NIQE model was trained on the LIVE image quality assessment database [19], which contains 29 reference images and 779 distorted images, each with 5 subjective quality scores. The distorted images were generated using a variety of distortion types, including compression, noise, and blur.

BRISQUE (Blind/Referenceless Image Spatial QUality Evaluator) is a no-reference image quality assessment algorithm [13]. It uses NSS to evaluate the quality of a distorted image without requiring a reference image for comparison. BRISQUE extracts 36-dimensional feature vectors from 96 non-overlapping blocks of the distorted image, which capture the statistical properties of the image. These features are then mapped onto a reduced-dimensional space using principal component analysis (PCA). The quality score is then calculated using a support vector regression (SVR) model that is trained on the LIVE IQA database [19].

Given the lack of reference images to assess the quality of the generated frames, traditional evaluation metrics like VIF and SSIM are not applicable to our problem. As a result, the efficacy of established no-reference image quality assessment (IQA) metrics, including NIQE and BRISQUE, was investigated.

The performance of NIQE and BRISQUE on real photo (a) and frames extracted from various T2V models was demonstrated in the results shown in Figure 3. These metrics evaluate images on a scale of 0 to 100, where higher scores indicate lower naturalness. It was observed that images (b) and (c) received the highest scores, indicating poor naturalness in the images. Furthermore, the presented examples show that non-natural images in (d) and (e) received better scores than the low-quality image of a dog in (b) and the image of an oil painting of a couple in (c), which still represent recognizable objects. Although NIQE scores showed slightly better results than BRISQUE, they were

still inadequate to fully differentiate between natural and non-natural images. Based on these findings, we opted to develop a new classifier to detect the naturalness of an image, recognizing that metrics such as NIQE and BRISQUE metrics are primarily concerned with the visual quality of generated videos rather than their naturalness.

(a) NIQE: 2.58 brisque: 23.85 (b) NIQE: 7.74 brisque: 54.41 (c) NIQE: 5.11 brisque: 39.86 (d) NIQE: 3.55 brisque: 15.75 (e) NIQE: 3.49 brisque: 22.88

Fig. 3: Image naturalness assessment with NIQE and BRISQUE

3 Our Proposed Metric

The main objective of our metric is to address the current limitations and concerns regarding the evaluation of T2V models, particularly related to image naturalness and modal biases. To address these challenges, we propose a novel ensemble metric that combines three different metrics to provide a more comprehensive evaluation. The metric workflow, as shown in Figure 4, can be divided into two parts.

The first part involves data generation, depicted in blue and yellow boxes on the left-hand side of the figure. Starting with an initial text prompt, we generate a video using a T2V model. Then, we use the generated video to produce a list of captions using BLIP-2 [6].

The second part involves the ensemble of three metrics, which starts with the Text Similarity Metric. This metric calculates the similarity score between the original text prompt and the BLIP-generated captions, ranging from 0 to 1. Next, we use the Naturalness Metric, a customised XGBoost classifier that takes the generated video as input and outputs a score ranging from 0 to 1.. Lastly, Average Precision is calculated using the objects in the video based on the original prompt and the BLIP-generated captions, with a possible range of 0 to 1.

In order to aggregate the evaluation metrics, a weighted average based on a linear regression(LR) model that was trained using manually rated videos is employed. This approach enables us to incorporate variations in each metric, as well as any potential biases or inconsistencies that could emerge from using a single metric. It is important to note that this technique has limitations, particularly in terms of the dataset size used to train the LR model and our own human biases. To address these limitations in future work, we aim to broaden the number of videos used for evaluation and to include outside human evaluation.

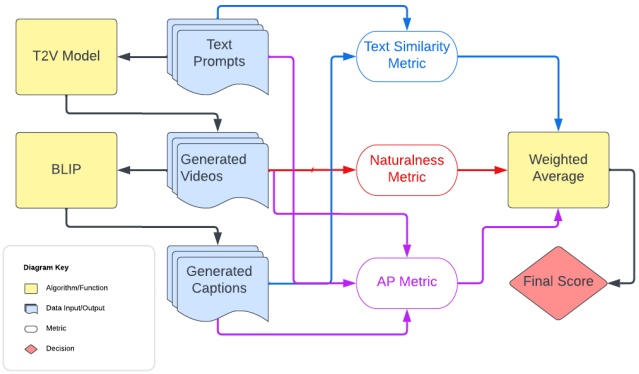


Fig. 4: Proposed Metric Ensemble

3.1 Evaluating Image Naturalness

Image Statistics used We have developed a customized classifier for video naturalness, aimed at differentiating between natural and non-natural content. To achieve this goal, we collected and analyzed several statistical measures from each video, including those outlined below.

1. **The texture score** measures the degree of uniformity in the texture of an image by calculating the variance of the image's gradient magnitude. This involves applying Sobel edge detection in the x and y direction to the image after it has been converted to grayscale and had a Gaussian blur applied to reduce noise. The magnitude of the gradient is then calculated, and the variance of this magnitude is used as the texture score. This is used to evaluate the naturalness of an image since natural images, such as landscapes or animal fur, tend to have more complex textures than synthetic images. $G(x,y) = \sqrt{G_x(x,y)^2 + G_y(x,y)^2}$ texture score=var(G(x,y))
2. **The sharpness score** function measures the amount of high-frequency content in an image, which is 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 image and the filtered image. $sharpness = \sqrt{mean((image - \text{filtered image})^2)}$
3. **The colour distribution score** is a measure of the uniformity of colour in an image. It is calculated by applying K-means clustering with K=2 to the A and B channels of the image's LAB representation. The cloud distribution score is then calculated as the proportion of pixels in the cluster with the lowest A channel value. It can help differentiate non-natural images, which may have more uniform or artificial colour distributions.
$$\text{cloud score} = \frac{\text{number of pixels in the cloud cluster}}{\text{total number of pixels}}$$
4. **The spectral score** measures the extent to which an image differs from the natural image statistics in the Fourier domain. The function calculates the mean and standard deviation of each colour channel of the image and

then computes the spectral score as the sum of standard deviations divided by the sum of means. spectral score $= \dfrac{\text{std}(R) + \text{std}(G) + \text{std}(B)}{\text{mean}(R) + \text{mean}(G) + \text{mean}(B)}$

5. **The entropy score** is being calculated using the Shannon entropy formula [2], which measures the level of randomness or disorder in pixel values. entropy $= -\sum_{i=0}^{255} p_i \log_2(p_i)$ Where p_i is the probability of intensity level i in the image, calculated as the histogram of the grey image divided by the total number of pixels. Natural images tend to have a higher degree of order and lower entropy than non-natural ones.
6. **The contrast score** measures the difference between the lightest and darkest parts of an image by calculating the standard deviation of the pixel intensities and then dividing it by the mean intensity.
7. **ORB (Oriented FAST and Rotated BRIEF)** is a feature detection algorithm [17] to compute statistics about the key points in an image, including the mean and standard deviation of the distances between key points and the mean and standard deviation of the lengths of the descriptors associated with the key points.
8. **The number and size of blobs** detected using the Laplacian of Gaussian (LoG) method [7]. Blobs are regions in an image with a relatively uniform intensity that stand out compared to the surrounding area.

No-Reference IQA

Despite their limited performance on certain images, NIQE and BRISQUE scores can still be useful in filtering out very noisy and disordered images. Therefore, we have included them in our evaluation process. To facilitate processing, the YUV444 video frame is reshaped from planar format 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 are calculated for the grayscale frame and YUV444 video frame's Y, U, and V channels separately. Separating the channels provides a better visual representation of the image [14] and allows for a more accurate evaluation of quality metrics like NIQE, which are more sensitive to variations in the chrominance channels.

To train a classifier we also calculated **Modified Inception Score(MIS)** for each video. We propose a modified version of the Inception Score (IS) metric. The MIS operates on a similar principle as described in 2.2, with the aim of assessing the quality of the generated videos by calculating the mean probability distribution of the frames in a video. However, we modified the metric in order to return a larger value if the mean probability distribution of the frames in a generated video has low entropy. Essentially if the Inception model assigns a greater probability to one particular class throughout the frames in a video, the MIS will produce a larger value. We achieved this by setting the marginal distribution to the uniform distribution.

Image Naturalness Classifier

After collecting all the video-representing data as described above, from 187 videos, including 92 natural videos and 95 non-natural videos, we approached the task as a binary classification problem. we approached the task as a binary classification problem. We manually assigned each video a label representing whether it was natural or not. Our approach involved training three classi-

fiers, AdaBoost, a Bagging classifier with a DecisionTree base and XGBoost. To optimize the performance of each classifier, we employed GridSearch. We evaluated the classifier's performance using the F1 score on the training, validation, and test sets. The XGBoost classifier performed the best on unseen data, demonstrating its superior ability to accurately classify natural and non-natural videos. The overall performance of the trained classifier on the train and test set is presented in Table 1.

Train Set		
Accuracy	F1 Score	Confusion Matrix
0.9677	0.9682	74 2
		3 76
Test Set		
Accuracy	F1 Score	Confusion Matrix
0.7500	0.7407	11 4
		3 10

Table 1: XGBoost Classifier Results on the training and test set

3.2 Natural Language Processing

The purpose of this section is to evaluate the textual semantic similarity between the generated video caption and the original caption. The process involves generating captions for each video frame using BLIP and calculating the similarity between each caption. By preprocessing the text, the textual data was made more manageable and easier to analyze, removing stop words and punctuation, adjectives and adverbs leading to better results in terms of similarity calculations and other text-based tasks. In our approach, we have decided to combine BERT and Cosine similarity ($cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|}$). Given the example presented in Figure 5 it can be seen that BERT tends to over-perform. BERT is designed to capture more nuanced and complex semantic relationships between sentences, whereas the cosine similarity metric only considers the surface-level similarity based on word overlap. By penalizing the BERT similarity score with the cosine similarity score, we can ensure that the combined similarity score Equation 1 reflects both the surface-level and deeper semantic similarities between the two sentences, thus providing a more accurate representation of their overall similarity. After conducting thorough analysis and running multiple experiments, we determined that the optimal ratio between BERT and cosine similarity would be 0.75 and 0.25, respectively.

$$\text{combined sim} = \begin{cases} 0.25(\text{cos sim}) + 0.75(\text{ BERT sim}), & if \text{cos sim} \neq 0 \\ 0.5(\text{BERT sim}), & otherwise \end{cases} \quad (1)$$

Given that some frames in the generated videos may exhibit significant distortions or may not contain recognizable objects, such as in the example pre-

Measure	Score
Cosine Similarity	0.00
BERT Similarity	0.45
Combined Similarity	0.22

(a) S1: The sunrise was beautiful over the ocean
S2: The bulldozer was loud and destroyed the building

Measure	Score
Cosine Similarity	0.288
BERT Similarity	0.765
Combined Similarity	0.65

(b) S1: A hot air balloon in the sky
S2: Balloon full of water exploding in extreme slow motion

Fig. 5: Similarity scores for two pairs of sentences

sented in Figure 6 where two frames do not include a dog. We opted to calculate the weighted textual similarity for generated video Weighted Similarity = $\frac{1}{n}\sum_{i=1}^{n} w_i \cdot \text{sim}_i$. The weights are assigned based on the frequency of each caption in the overall list of generated captions.

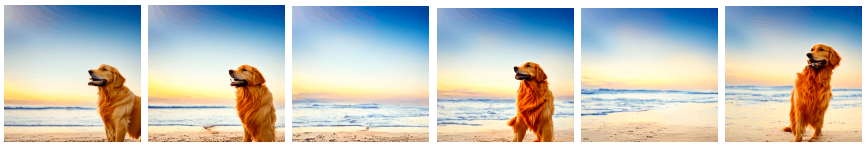


Fig. 6: A golden retriever eating ice cream on a beautiful tropical beach at sunset

3.3 Semantic Matching

The CLIPSim metric is commonly used to evaluate semantic matching between input prompts and generated videos. This metric produces a score between 0 and 1, where a score closer to 1 indicates a successful semantic match. However, we have observed flaws in CLIPSim, which may over-fit and return higher scores for models that utilize CLIP as part of their architecture such as Make-a-Video[20] and Aphantasia, regardless of their video generation abilities. To address this issue, we propose a two-step semantic matching metric based on bounding-box object detection comparison between the original text prompt and the BLIP-2[6] generated caption.
Methodology: We collected 35 text prompts from Make-a-Video and Imagen Video[4] papers to generate a total of 175 videos using the T2V models Tune-a-Video, VideoCrafter, Aphantasia, VideoFusion and Text-to-Video Synthesis. Each frame of the generated videos was passed through BLIP to obtain a list of captions for each frame. We used Grounding DINO (GDINO)[8] to generate bounding boxes for each frame using the original text prompt and the BLIP caption. We extracted relevant information from both prompts and passed it along with the frame into GDINO. We present an example of this process in Figure 8, the model ran GDINO on the frame of "Twins eating ice cream

on top of Eiffel tower" twice, on the first pass through the model used the extracted phrases "twins" and "Eiffel tower" from the original caption and in the second pass through it took the extracted BLIP phrases "dolls" and "tower" as presented in Figure 8. The model detected six bounding boxes, two for the "twins"/"dolls" seen in red and green and one for "Eiffel Tower" (Figure 8a) and "tower"(Figure 8b) in yellow. The model then calculates the IoU of the bounding boxes with the original caption treated as the ground truth. This process was repeated for each frame in a video. Finally, the bounding boxes and phrases of each frame were used to calculate the Average Precision (AP) value for the video. AP is calculated as explained by A. Anwar[5].
To provide a clear explanation of our method, we present Figure 7. The first step of our approach involves collecting a list of prompts, followed by selecting the T2V models to generate videos from the text prompts. The generated videos are then fed into the BLIP-2 model to generate captions. The captions, along with the original text prompts and generated videos, are used to generate bounding boxes for each frame in the videos. The second step of our approach uses the outputted bounding boxes and phrases to calculate the IoU of the bounding boxes and then the AP for the videos. The final score is the mean AP value for each model. Our proposed two-step semantic matching metric improves on the CLIPSim metric by addressing the over-fitting issues that can occur with models that use CLIP.

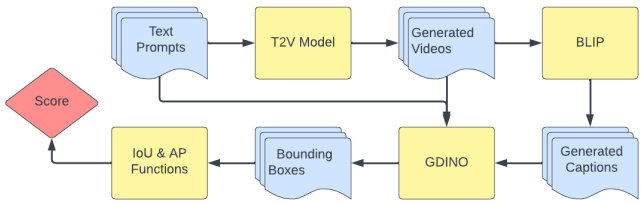


Fig. 7: Average Precision Architecture

4 Evaluation

We computed the three metrics, naturalness score, text similarity score, and AP, for each of the 35 videos using four different models. To obtain the weights for each of the customized scores, we trained a Linear Regression model. The results of the Linear Regression model are presented below, and it can be seen that the model performs poorly on a per-video basis, with an R2 score of only about 0.33 which means that the model is able to explain only 33% of the variability in the target variable. This suggests that the model has limited predictive power and may not be suitable for accurately predicting the human

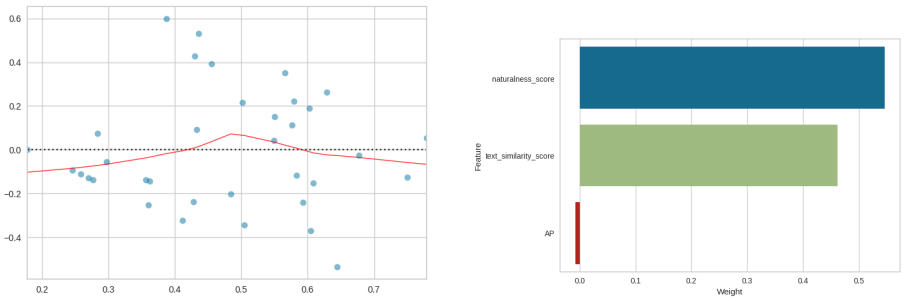
[5] Intersect over Union Explained

(a) Ground Truth Bounding Boxes (b) BLIP Bounding Boxes

Fig. 8: GDINO Example

evaluation score.

In this section, we present the evaluation of our proposed metric for T2V



(a) Residuals Plot for Linear Regression Model Performance

(b) Weights assigned to each feature by the linear regression model.

models, as shown in Table 2. Our evaluation approach involved analyzing the performance of each T2V model on our three novel metrics separately and also on the weighted score. To ensure a fair comparison, we used BLIPSim (CLIPSim alternative) to evaluate all the generated videos. Additionally, we manually ranked each video on a scale of 1 to 5, which represents the human evaluation. These human evaluation scores were used to compare our proposed metric to BLIPSim, and we found that our metric more closely correlates with the mean human evaluation of each model than BLIPSim does. Specifically, our metric returned a correlation coefficient of 0.98, while BLIPSim returned 0.61. We acknowledge that our results may not be conclusive due to the limited number of videos used and our own biases that may be present in the human evaluation scores. Nonetheless, our proposed metric represents a significant improvement in evaluating the performance of a T2V over a number of videos but performance poorly on individual videos.

Score	T2VSynthesis	VideoCrafter	Tune-a-Video	Aphantasia
Naturalness↑	0.66	**0.67**	0.65	0.36
Text Similarity↑	0.58	**0.62**	0.61	0.47
AP IoU↑	**0.45**	0.35	0.29	0.16
Weighted Score↑	0.62	**0.65**	0.63	0.40
BLIPSim↑	**0.66**	0.57	0.57	0.54
Human Evaluation↑	0.59	**0.62**	0.5	0.09

Table 2: Average Performance Across All Models

4.1 Results

In light of the unsatisfactory performance of the Linear Regression classifier, we also provide an individual video evaluation results. This section presents a comprehensive overview of the outcomes obtained from our customized evaluation metric. For reference, we present the output of four T2V models using two different text prompts as in Figure 10.

A small domesticated carnivorous mammal with soft fur, a short snout, and retractable claws

T2V Synthesis VideoCrafter Tune-a-Video Aphantasia

A happy elephant wearing a birthday hat walking under the sea

Fig. 10: T2V Models Output Examples

Based on the results presented in Table 3, it can be concluded that the performance of the T2V models varied greatly depending on the input prompt. The three evaluation metrics used in this study - naturalness, text similarity, and AP - complement each other in assessing the overall performance of Text-to-Video models. Naturalness measures the level of human perception of the generated video, text similarity assesses how well the generated video corresponds to the input prompt, and AP measures how accurately the generated video reflects the input text in terms of visual content.

Score	**T2VSynthesis**	**VideoCrafter**	**Tune-a-Video**	**Aphantasia**
Naturalness↑	0.35	0.57	**0.94**	0.52
Text Similarity↑	**0.29**	0.28	0.22	0.23
AP IoU↑	0.0	0.0	0.0	0.0

"A small domesticated carnivorous mammal with soft fur, a short snout,and retractable claws"

Naturalness↑	0.74	**0.83**	0.67	0.29
Text Similarity↑	**0.83**	0.73	0.74	0.51
AP IoU↑	0.0625	0.4	**0.85**	0.06

"A happy elephant wearing a birthday hat walking under the sea"

Table 3: Individual Results for T2V Models on Given Prompts

Based on the given results, it can be observed that all models have demonstrated relatively higher naturalness scores compared to the Aphantasia model. This implies that the videos generated by Aphantasia are less human-interpretable. The Naturalness score provided valuable insights into the realism of the generated videos, as it was observed to perform well in distinguishing more cartoonish frames from more realistic ones. For instance, in the example provided in Table 3a, the score of 0.67 assigned to a frame generated by Tune-a-Video was relatively lower due to its cartoonish appearance, whereas in another example Table 3b, the same model was assigned a score of 0.94 due to its realistic appearance when compared to frames generated by other models for a given prompt. These findings suggest that the XGB Classifier utilized for evaluating the visual realism of generated videos is producing highly accurate and precise results.

The Text Similarity and AP IoU metrics provide valuable insights into how well the generated video aligns with the text prompt used to generate the video. These metrics help evaluate the model's ability to accurately capture the intended meaning of the input text and generate a video that is semantically meaningful and coherent. However, it was found that longer and more complicated captions tended to yield lower performance results, even when using preprocessing to remove extraneous information. It proves that the models were more successful in generating videos that had a higher degree of overlap with the training data. In particular, simpler and more straightforward captions tended to yield better results. These findings suggest that there is still a need to improve the generalization ability of the models to generate more diverse and complex videos from a wider range of prompts.

5 Conclusion

Our work introduces the development of a novel technique for evaluating text-to-video (T2V) and a comprehensive discussion on commonly employed evaluation metrics in this field. In this paper, we aimed to address the limitations of

prevailing evaluation methods and presented a novel evaluation approach that focuses on assessing the naturalness of visual content as well as the degree of semantic correspondence between the video and its corresponding text. To enhance the accuracy and efficacy of our evaluation metric, we intend to augment our dataset with a larger and more diverse sample of videos. Additionally, we plan to conduct further experiments by soliciting feedback from a broader audience, thereby expanding the scope of our human evaluation process. Through these measures, we aim to achieve more robust and comprehensive results, ultimately improving the overall performance of our evaluation metric. Our primary contribution to the field of T2V is the proposed evaluation technique, which enhances the existing evaluation frameworks and provides a more reliable measure of the quality of generated videos. This study can contribute to the advancement of T2V technology by providing a more comprehensive and effective evaluation methodology, thus improving the quality and relevance of generated videos.

References

1. Barratt, S., Sharma, R.: A note on the inception score. arXiv preprint arXiv:1801.01973 (2018)
2. Cover, T.M., Thomas, J.A.: Elements of information theory second edition solutions to problems. Internet Access pp. 19–20 (2006)
3. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. Advances in neural information processing systems **30** (2017)
4. Ho, J., Salimans, T., Gritsenko, A., Chan, W., Norouzi, M., Fleet, D.J.: Video diffusion models. arXiv preprint arXiv:2204.03458 (2022)
5. Khachatryan, L., Movsisyan, A., Tadevosyan, V., Henschel, R., Wang, Z., Navasardyan, S., Shi, H.: Text2video-zero: Text-to-image diffusion models are zero-shot video generators. arXiv preprint arXiv:2303.13439 (2023)
6. Li, J., Li, D., Savarese, S., Hoi, S.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. arXiv preprint arXiv:2301.12597 (2023)
7. Lindeberg, T.: Feature detection with automatic scale selection. International journal of computer vision **30**(2), 79–116 (1998)
8. Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Li, C., Yang, J., Su, H., Zhu, J., Zhang, L.: Grounding dino: Marrying dino with grounded pre-training for open-set object detection (2023)
9. Luo, Z., Chen, D., Zhang, Y., Huang, Y., Wang, L., Shen, Y., Zhao, D., Zhou, J., Tan, T.: Decomposed diffusion models for high-quality video generation. arXiv preprint arXiv:2303.08320 (2023)
10. Mack, D.: A simple explanation of the inception score. https://medium.com/octavian-ai/a-simple-explanation-of-the-inception-score-372dff6a8c7a (Mar 2019)
11. MathWorks: Train and use a no-reference quality assessment model. https://www.mathworks.com/help/images/train-and-use-a-no-reference-quality-assessment-model.html (2021)
12. Mittal, A., Moorthy, A.K., Bovik, A.C.: No-reference image quality assessment in the spatial domain. IEEE Transactions on image processing **21**(12), 4695–4708 (2012)

13. Mittal, A., Soundararajan, R., Bovik, A.C.: Making a "completely blind" image quality analyzer. IEEE Signal processing letters **20**(3), 209–212 (2012)
14. Podpora, M., Korbas, G.P., Kawala-Janik, A.: Yuv vs rgb-choosing a color space for human-machine interaction. In: FedCSIS (Position Papers). pp. 29–34. Citeseer (2014)
15. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning. pp. 8748–8763. PMLR (2021)
16. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10684–10695 (2022)
17. Rublee, E., Rabaud, V., Konolige, K., Bradski, G.: Orb: An efficient alternative to sift or surf. In: 2011 International conference on computer vision. pp. 2564–2571. Ieee (2011)
18. Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., Chen, X.: Improved techniques for training gans. Advances in neural information processing systems **29** (2016)
19. Sheikh, H.R., Bovik, A.C.: Image information and visual quality. IEEE Transactions on image processing **15**(2), 430–444 (2006)
20. Singer, U., Polyak, A., Hayes, T., Yin, X., An, J., Zhang, S., Hu, Q., Yang, H., Ashual, O., Gafni, O., et al.: Make-a-video: Text-to-video generation without text-video data. arXiv preprint arXiv:2209.14792 (2022)
21. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2818–2826 (2016)
22. Unterthiner, T., van Steenkiste, S., Kurach, K., Marinier, R., Michalski, M., Gelly, S.: Fvd: A new metric for video generation (2019)
23. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE transactions on image processing **13**(4), 600–612 (2004)
24. Wu, C., Huang, L., Zhang, Q., Li, B., Ji, L., Yang, F., Sapiro, G., Duan, N.: Godiva: Generating open-domain videos from natural descriptions. arXiv preprint arXiv:2104.14806 (2021)
25. Wu, J.Z., Ge, Y., Wang, X., Lei, W., Gu, Y., Hsu, W., Shan, Y., Qie, X., Shou, M.Z.: Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. arXiv preprint arXiv:2212.11565 (2022)

6 Appendix

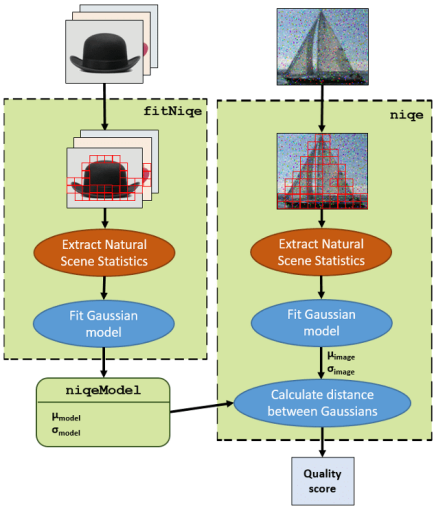


Fig. 11: NIQE workflow [11]

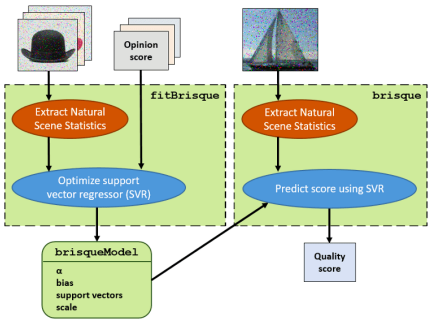


Fig. 12: BRISQUE workflow [11]