# The Lost Melody: Empirical Observations on Text-to-Video Generation From A Storytelling Perspective

Andrew Shin Keio University

shin@inl.ics.keio.ac.jp

Yusuke Mori The University of Tokyo

mori@mi.t.u-tokyo.ac.jp

Kunitake Kaneko Keio University

kaneko@dmc.keio.ac.jp











Figure 1. Scenes from the video "The Lost Melody" generated from a short story.

# **Abstract**

Text-to-video generation task has witnessed a notable progress, with the generated outcomes reflecting the text prompts with high fidelity and impressive visual qualities. However, current text-to-video generation models are invariably focused on conveying the visual elements of a single scene, and have so far been indifferent to another important potential of the medium, namely a storytelling. In this paper, we examine text-to-video generation from a storytelling perspective, which has been hardly investigated, and make empirical remarks that spotlight the limitations of current text-to-video generation scheme. We also propose an evaluation framework for storytelling aspects of videos, and discuss the potential future directions.

## 1. Introduction

Text-to-video generation has gained attention with its increasingly better controllability, visual aesthetics, and potential for a wide range of applications, including scene generation and video editing. Creating a novel video whose duration ranges from a few seconds to a few minutes can now be done within a few minutes by simply providing text prompts, and a plethora of videos generated from text have appeared on social network services and streaming services, many of which demonstrate high-quality visuals.

However, current text-to-video generation models almost invariably focus on generating visual components of a

single scene or movement, while mostly disregarding other modalities including speech or text. As such, current text-to-video generation shares a high resemblance with movies from the silent film era, in which dialogues were either conveyed through text or completely forgone, with background music inserted to deliver the atmosphere. The difference lies in that, while most silent films strove to tell a concrete story with structured plots in spite of their lack of auditory tools, current text-to-video generation models tend to focus on presenting an array of imagery centered around the prompt, rather than storytelling. It thus becomes our central question whether contemporary text-to-video generation can also mimic the storytelling aspects of silent films that share similar limitations.

In this paper, we examine current text-to-video generation models from a perspective of storytelling. We first examine generating videos from short stories generated by a large language model, one of which is entitled "The Lost Melody" (Figure 1), for which we make frequent references throughout the paper in order to illustrate important points¹. We also examine generating videos from script version of each short story, in order to illustrate a striking performance drop stemming from asymmetry in training data. We further examine generating videos from captions that are aimed to perform storytelling for existing videos, in order to contrast our generated results with actual videos and captions aimed for storytelling.

<sup>&</sup>lt;sup>1</sup>Sample videos are publicly available on YouTube, and can be retrieved with the title of this paper as keyword.

Subsequently, We propose and conduct an extensive evaluation to properly assess the videos from various aspects, including components of a story, visual qualities, and how they correlate to each other. We also propose a novel evaluation framework T2Vid2T, where captions are generated from videos and are compared to input text prompts or ground truths, in a cyclical manner. Our evaluation protocols utilize both automatic and human evaluation metrics extensively, in order to account for the videos both qualitatively and quantitatively.

Upon discussion and analysis of the generation and evaluation carried out as described above, we further make empirical remarks as to the limitations that make it difficult to generate videos that successfully tell a story, and discuss potential future directions that we believe would help critically refine the performance.

Note that, since our primary focus is on storytelling, we deliberately disregard visual artifacts (which frequently appear throughout our experiments) in most cases, although our experiments suggest that they may also play a nonnegligible role in storytelling. We also disregard other shooting or editing techniques frequently used in modern videos, such as changing shots and angles for more dramatic impact.

Our primary contributions can be summarized as following:

- We examine contemporary text-to-video generation scheme from a hardly investigated storytelling perspective
- In doing so, we propose a novel evaluation framework for videos generated from text
- In our extensive discussion and analysis, we further suggest potential future directions for enhancing textto-video generation in terms of reflecting storytelling aspects.

#### 2. Related Work

**Text-to-Video Generation**: Text-to-video has become an actively deployed research topic. Frequently employed architectures include diffusion model [15] and transformer [45], and also models that enable learning of common embedding for text and visual inputs, such as CLIP [34].

Make-A-Video [40], built on top of text-to-image generation model DALL·E 2 [36], which in turn is based on CLIP, generates videos using spatio-temporal convolution and frame interpolation networks. Temporal corrections are made using a U-net-style diffusion model, and frame interpolation network fills in the gaps between frames generated by spatio-temporal decoder, resulting in a smoothly moving video. CogVideo [16], also based on text-to-image generation model CogView2 [9], utilizes VQVAE [44] to convert each frame of a video into image tokens. Since applying

the same frame rate to all videos can lead to mismatches between the content of the videos and the text, and consecutive frames tend to have very similar content, making it challenging to learn long-term dependencies with a fixed frame rate, CogVideo inserts a token indicating the frame rate into the text and samples frames at the specified frame rate. Imagen Video [14] utilizes a cascade of spatio-temporal superresolution, generating a total of 128 frames with a resolution of 1280×768 at a frame rate of 24fps. The advantage of such cascade model is that it can independently learn each super-resolution diffusion model. The text encoder uses a pre-trained T5 [35] model with fixed weights. Imagen Video is particularly noteworthy for its ability to represent text within videos, a task that was challenging for traditional video generation models.

Sora [3] is a diffusion model using transformer architecture, and enables generation of videos up to 1 minute that demonstrate high fidelity to input prompts with highly plausible graphics. Recaptioning technique from DALL·E 3 [31] has been said to have helped generating videos with high fidelity to users' text input prompts.

While many text-to-video generation models do not allow for public examination and explicitly state the limitations, it is generally fair to say that most of the text-to-video generation models above presuppose that the input textual prompt aims to describe a single scene or motion, such that, upon receiving prompts that contain multiple scenes as necessary for a storytelling, the models usually end up reflecting only a limited portion of the input, failing to generate results that successfully perform storytelling. Its causes may be attributed to various factors including limitations on input length, duration of output, and training data.

Visual Storytelling: As a more advanced task than conventional image captioning task, visual storytelling [19] was proposed to generate a story based on a sequence of images. In contrast, [23] proposed the story visualization task, attempting to generate a sequence of images from a story. Examining video generation from stories is a natural extension of the line of works above that attempt to bridge the gap between vision and storytelling. Recent works have tackled the task of consistent multi-scene text-to-video generation [25,27], but their focus is on multi-scene generation from a single prompt, rather than a wider scope of storytelling.

**Story Evaluation**: Automatic evaluation of stories is an essential research topic for tasks related to stories. However, in textual story evaluation, it has been pointed out that existing metrics correlate poorly with human evaluation [13]. In order to overcome such low correlation, novel evaluation metrics including UNION [12] and StoryER [6] have been proposed and demonstrated improved correlations with human perception.

Such limitation of conventional evaluation metrics has

also been pointed out in visual storytelling [17, 48]. [18] proposed a metric for visual storytelling from three perspectives; namely *relevance*, *coherence*, and *expressiveness*. Inspired by [18], [47] proposed the similar three perspectives to evaluate visual storytelling; *visual grounding*, *coherence*, and *non-redundancy*. To assess story quality, [26] used story-specific metrics in addition to lexical-matching metrics.

# 3. Components of a Story

While various definitions exist as to what constitutes a story, common elements include character, plot, setting, conflict, theme, and point of view. We briefly describe each component and discuss the challenges it poses on videos generated from text, which will also be examined explicitly later in the paper.

Character is the person in the story, around which the story evolves. In this simple sense, we can say that videos generated from text also frequently contain characters. However, there is rarely a single character in a story, and multiple characters including a protagonist, the main character, and an antagonist, an opposition to the protagonist, are typically present. Furthermore, character can also refer to the qualities of a person. Direct characterization, where the qualities are directly presented to the audience, such as their appearances, can be easily manipulated in videos generated from text, and certain aspects of their personalities can often be specified by adjusting facial expressions. Indirect characterization, however, where the qualities are formed through the person's statements or behaviors, inevitably necessitates a development of a plot.

Setting refers to time and location in which the story takes place. The degree of specification required varies depending on the story, and videos generated from text inherently provide direct visual clues for the setting. Also, setting may contain the social conditions under which the character is placed, which is presented less directly and requires context. Although not entirely impossible, it is certainly more challenging for videos generated from text to accurately present.

Plot refers to the sequence of events that occur throughout the story, and develops through multiple stages including introduction, rising action, climax, falling action, and resolution. While it is certainly possible in theory to demonstrate these multiple stages solely with visual elements, as was the case with silent movies, it turns out to be difficult with current text-to-video generation, as we will see later.

Conflict, often considered to be an element of plot, refers to the primary opposition that the character undergoes, and may be internal, *i.e.* versus self, or external, *e.g.* with other characters or circumstances. In any case, insomuch as the plot is nearly absent, it is evident that conflict is also difficult to convey. Theme refers to the central message of

the story, and may not always be directly expressed. While visual elements often strongly contribute to the highlighting of a theme, it is typically developed through the plot, e.g. the outcomes of a character's behaviour. As such, it is highly challenging for current text-to-video generation to effectively express a theme. Point of view in a story of text format refers to the narrator; first person if the narrator is a character within the story, and third person if the narrator is not. Point of view takes a slightly different meaning when applied to videos. Implementing a truly first person's point of view would be to use views seen from the character's eyes, and while such shot is rare in commercial films and television shows, it can be generated by specifying it in input prompts. Third person's point of view, where the audience oversees what is happening as an observer, is in most cases the default point of view for videos generated from text, unless specified otherwise.

In addition, other elements may be included, such as style or tone of the story, and symbolism. The former can be conveyed through the visual atmosphere of the video, which can also be controlled with specifications in input prompts. The latter is more of a derivative of a plot, and thus can hardly exist in the absence of a plot.

In summary, current text-to-video generation, while inherently containing some elements of a story to varying extents, may struggle to fully attain the indispensable elements that constitute a story, as we will empirically observe later. In this paper, we primarily focus on character, setting, and plot, as other components can be considered subcomponents of one of these, and without accounting for the those three components, reflecting the remaining components will be simply out of question.

## 4. Video Generation

We now describe our workflow of generating videos for stories. We examine generation from three distinct types of text prompts; short stories generated by LLM, scripts generated by LLM, and captions that describe existing videos.

## **4.1. Generation from Short Story**

We used ChatGPT [30] with GPT3.5 to generate a short story from which to generate a video. The generated short story consists of multiple scenes. As discussed in Section 2, with current text-to-video generation, it is nearly infeasible to generate a video for the entire story from a single round of generation, in terms of both maximum token length acceptable and the model's ability to compose multiple scenes. As such, we need to convert the generated short story into a sequence of prompts, where each prompt corresponds to a scene. Concatenating the generated video for each prompt will be the resulting video of the short story.

Note that each scene is generated independently of each other, without any explicit adherence to previously gener-

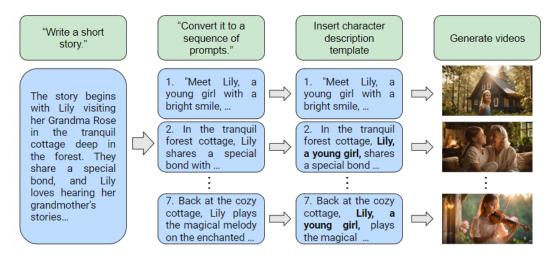


Figure 2. Overall workflow of generating videos from a short story generated by a large language model.

ated outcomes. As such, each prompt for video generation model needs to specify the characters and the setting repeatedly. Otherwise, the context is easily lost and the generation results in incoherent videos. An alternative would be to generate the videos conditioned on previously generated outcomes. We did examine conditioning video generation on an image from previously generated results, but the result turned out to be unreliable as we will further discuss in Section 6.

Since each scene is generated independently, we also insert character description for proper nouns repetitively in each prompt. For example, "Lily shares a special bond with ..." is replaced with "Lily, a young girl, shares a special bond with...," and so forth for each prompt. Otherwise, apart from the first appearance of the character, none of the visual descriptions will be available in subsequent scenes, resulting in immense inconsistency, as illustrated in Figure 3. As we will re-visit in Section 6, a scheme to integrate previous context, frequently used in LLMs, may help achieve visual coherence more easily.

Finally, video is generated for each prompt with Gen-2 [10], after which the generated videos are temporally concatenated. In order to examine how the comprehensibility of the story changes with linguistic aid, we generate speech from each prompt using OpenVoice [33]. We adjuste the playing speed of each scene in order to match the duration of narration. The same duration is applied to the version without narration, as the varying duration can affect the viewer's understanding of the story. For aesthetic purpose, we also generate the background music with Stable Audio [11] using the scene description as the prompt, which was applied to all of our videos in an identical manner.

Figure 2 shows the overall workflow of generating videos for a short story created by an LLM. See Supplemental Material for actual prompts we used and the generated outcomes.

"In the tranquil forest cottage, Lily shares a special bond with Grandma Rose. Lily cherishes hearing her grandmother's enchanting stories."



Figure 3. Unless each prompt contains a sufficient amount of details, the model generates incoherent results.

#### 4.2. Generation from Script

Text-to-image generation or text-to-video generation have mostly evolved around descriptive text prompts, most of which are concerned with vividly conveying the visual components to be generated. Yet, humans can also visualize the scene from dialogue, which is in fact a strong medium for conveying a narrative along with novel. In fact, dramas or plays written in script format have longer history than novels, and have been around for millennia. In order to examine to what extent video generation can handle storytelling given text prompt in script format, we generate a script version of the short story generated as described above. We use LLM to generate a sequence of prompts for video generation model, explicitly requiring the dialogues spoken by characters, with non-dialogue parts for indicating setting or transitions. As before, each scene is generated independently, so proper nouns in each dialogue are supplemented with description for each prompt.

Table 1. FVD and inception score on each type of videos. Lower FVD and higher inception score indicate better results.

Model	FVD(↓)	Inception Score(↑)
Generated (story)	51.53	11.01
Generated (script)	75.34	10.25
Generated (caption)	46.06	12.22
Real videos	6.54	32.18

### 4.3. Generation from Captions

In order to make a comparison with how story is perceived from real videos, we sampled a subset from Video Storytelling dataset [22], whose ground truth captions were used for video generation. While other datasets that examined visual storytelling exist, such as SIND [19], most of them contain a sequence of images rather than videos, and are thus not apt for our purpose. Since videos in Video Storytelling dataset usually span several minutes, which are lengthier than our generated videos, we either used the first 5 captions, or the first 100 seconds of the video, selecting whichever one results in shorter duration.

# 5. Experiments

#### 5.1. Evaluation

As shown in Section 2, while there have been evaluation metrics proposed for textual storytelling or video captioning, evaluation of videos from a storytelling perspective has hardly been examined, to the best of our knowledge. In this section, we propose and conduct a wide range of evaluation metrics to assess storytelling ability of videos both qualitatively and quantitatively, involving both automatic and human evaluations.

We first evaluate the visual quality of the generated videos with FVD (Fréchet video distance) [43] and inception score (IS) [39]. For FVD, following [7], we used I3D [5] trained on Kinetics-600 [4] to compute the real video statistics. For inception score, following [38], we used pretrained C3D [42], first trained on Sports-1M [20] and finetuned on UCF101 [41]. We extracted 10 frames per second from the target videos and report the average values. While these metrics are not directly relevant to storytelling *per se*, we will later use them to examine how they correlate to the audience's perception of the story.

We also propose a novel cyclical evaluation framework, namely T2Vid2T (text-to-video-to-text), where we automatically generate captions that describe the generated video, and compare it against the original text prompt that was used to generate the video. We used a fine-tuned version of TimeSFormer [2] for video captioning to generate captions from each scene. Popular evaluation metrics including BLEU [32], METEOR [8], ROUGE [24], CIDEr [46], and SPICE [1] were employed for comparison against

text prompt. Note that we replace the proper nouns in text prompt with general descriptive phrase, e.g. "Lily" with "a young girl", as the proper noun is out of scope with video captioning framework. For Video Storytelling, we disregarded the parts where ground truth captions were not provided, and extracted only the frames that were accompanied by captions. Figure 4 illustrates the workflow of T2Vid2T using Video Storytelling as the reference.

We further conducted human evaluation, where we asked the workers to rate the video in terms of the components of the story, namely character, setting, plot, as outlined in Section 3, out of 1 to 5 scale. Each component of the story was rated according to three distinct aspects, namely *expressiveness*, *relevance*, and *coherence*, as suggested by [18], yielding a 3 × 3 evaluation matrix per video. *Expressiveness* refers to richness and diversity of expression styles. *Relevance* refers to the extent to which the video reflects the text prompt in terms of the story component of concern. *Coherence* refers to the extent to which the consistency of respective component is retained throughout the video. In addition, a worker was asked to rate the video in terms of overall comprehensibility of the story.

We used Amazon Mechanical Turk for our experiment. We presented the workers with videos generated from a short story, a script version of the story, and a caption from Video Storytelling, along with an actual video to which the caption belongs. Each generated video is presented in two variations; the generated video as is, and the generated video with narration added to it. Note that human evaluation was performed with 2 conditions, where the worker was either explicitly asked to ignore the visual artifacts, or to take visual artifacts into consideration, in order to examine how they contribute to perception of visual storytelling.

Finally, we conducted a mixture of human and automatic evaluation, where the worker was asked to write a summary of the video, with one sentence per scene, and the resulting summary was compared against the input textual prompt using automatic evaluation metrics.

#### 5.2. Results

Table 1 shows FVD and inception score for each type of videos. Real videos from Video Storytelling outperform other generated videos by far. While such result is highly predictable, it is particularly of our interest that higher visual qualities also correlate to perception of story, as will be shown in other results. Also, videos generated from captions for Video Storytelling displayed better results than videos generated from short stories or scripts, suggesting that current text-to-video generation models are more compatible with factual descriptions.

Table 2 shows the results of T2Vid2T. Note that, input prompts were used as ground truths for results other than Video Storytelling. Video captioning is still widely consid-

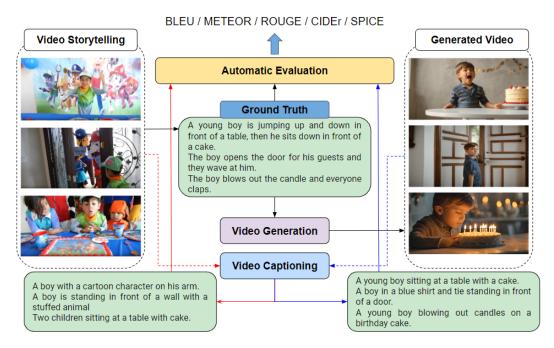


Figure 4. Evaluation workflow for T2Vid2T.
Table 2. Results of T2Vid2T

Model	BLEU-1	METEOR	ROUGE-L	CIDEr-D	SPICE
Generated (story)	0.0298	0.0343	0.1525	0.0275	0.0877
Generated (script)	0.0298	0.0258	0.1325	0.0011	0.0440
Generated (caption)	0.1045	0.0635	0.1244	0.0615	0.0952
Real videos	0.2630	0.1683	0.2572	0.2604	0.3519

ered a relatively difficult task, as can also be seen by notably low scores throughout all videos. Setting aside the real videos, which performed far better as predicted, it is notable that videos generated from stories and scripts performed substantially poorly, even when compared to videos generated from captions. This suggests that video captioning model struggles even to match a few tokens with the input stories or scripts, hinting at a severe imbalance in the training corpora in terms of textual styles and formats.

Table 3 summarizes the results from human evaluation in terms of how well the video reflects components of a story. While trends from previous results generally continue, adding linguistic elements via narration improved the overall scores. It is notable that in some categories, videos with narrations received lower scores than videos without narrations. We conjecture that this is due to mismatch between narration and the generated scene. For example, texto-video generation frequently struggles with multiple characters, and easily ends up generating incoherent characters or skipping the generation of some characters present in input prompt. Such mismatch is likely to exacerbate, rather than aid, the perception of story. Asking the workers to take visual artifacts into consideration resulted in performance

drop, suggesting that visual qualities also play a role in perception of stories. In fact, our investigation showed that visual artifacts also had an influence on automatic evaluations by incorrectly classifying the objects, e.g. violin to guitar, etc. Table 4 shows the results of running automatic evaluation metrics on human-written summary upon watching the videos, with input prompts as the ground truths. Its overall trend is coherent with Table 2, suggesting that the low performance of videos generated from stories and scripts is not solely attributed to problems with video captioning models, but is also highly relevant to limitations with textto-video generation models as well. Table 5 shows example human-written summary for each type of videos. Note that, for "The Lost Melody", later parts of the plot were often completely misunderstood by the workers. This reinforces our claims and experiment results that current text-to-video generation models are yet to generate convincing videos for storytelling. Figure 5 visualizes how the plot for a scene was misunderstood by the worker.

#### 6. Discussion

In this section, we review some of the notable drawbacks of current text-to-video generation models in terms of story-

Table 3. Ratings for generated videos in terms of the components of a story. Numbers in the parenthesis indicate the ratings when the workers were asked to take visual artifacts into consideration as well.

Model	Character		Setting		Plot			Overall		
Woder	Exp.	Rel.	Coh.	Exp.	Rel.	Coh.	Exp.	Rel.	Coh.	Comprehensibility
Gen. (story)	2.5(2.2)	2.0(1.7)	2.8(2.6)	3.0(2.5)	3.5(3.2)	2.8(2.6)	3.2(2.8)	3.2(3.1)	2.9(2.4)	2.7(2.0)
Gen. (story) (narration)	2.7(2.2)	1.8(1.7)	2.9(2.8)	3.3(2.9)	3.7(3.2)	3.0(2.5)	3.4(3.0)	3.0(3.0)	3.0(2.6)	3.2(2.5)
Gen. (script)	2.8(2.5)	1.7(1.6)	2.1(2.0)	2.8(2.6)	2.5(2.2)	1.5(1.3)	1.9(1.8)	2.0(1.9)	2.2(1.8)	2.1(1.9)
Gen. (script) (narration)	2.8(2.6)	1.4(1.2)	2.0(1.5)	2.8(2.4)	2.6(2.4)	1.7(1.6)	2.0(1.8)	2.2(2.2)	2.5(2.2)	2.5(2.2)
Gen. (caption)	3.5(3.2)	3.8(3.4)	3.6(3.3)	2.5(2.4)	3.2(2.7)	3.0(2.7)	3.5(3.0)	3.9(3.6)	3.7(3.3)	3.8(3.6)
Gen. (caption) (narration)	3.7(3.7)	4.0(3.8)	3.8(3.7)	2.8(2.4)	3.2(3.0)	3.1(2.7)	3.7(3.7)	4.2(3.9)	3.7(3.7)	4.0(4.0)
Real videos	4.6	4.7	4.8	4.5	4.2	4.0	3.8	4.8	4.9	4.8

Table 4. Results of automatic evaluations on human-written summary.

Model	BLEU-1	METEOR	ROUGE-L	CIDEr-D	SPICE
Generated (story)	0.0444	0.0595	0.1614	0.0038	0.1607
Generated (script)	0.0348	0.0537	0.1570	0.0036	0.1326
Generated (caption)	0.2467	0.1499	0.3377	0.7429	0.2889
Real videos	0.3096	0.2299	0.5137	1.6115	0.3060



Figure 5. Example of human-written summary that is substantially different from input.

telling as observed in our experiments, and discuss potential future directions that would enhance the performance.

Video generation with visual reference: We generated videos for the story simply by concatenating multiple generated videos for its constituents. Such workflow is bound to incur inconsistency in visual aspects, as we have witnessed in our experiments. An alternative would be to condition the text-to-video generation on a reference visual, from which to learn how the characters should look like, and reflect it in the generated videos. Indeed, a story generation scheme with additional source frame has been proposed for text-toimage generation [29], and some of the text-to-video generation models also enable both image and text inputs. However, it must be noted that, in storytelling perspective, the purpose of conditioning the generation on image and text is fairly different from current editing schemes based on image and text inputs. In text-based image editing, for example, input images themselves are targets to be directly

modified via directions in the text prompt. A direct extension of this to video generation, which is currently deployed by some of the text-to-video generation models, also entails an assumption that input images are targets to be directly modified in the generated video, e.g. generating motions on the image as specified by input prompt. As such, when confronted with text inputs that do not imply a direct editing of the input image, the resultant video utterly fails to reflect any component of the input prompt, and merely ends up making subtle motions on the input image (see Figure 6). In storytelling perspective, image or video inputs conditioned for the generation are frequently not the targets to be directly modified, but are references that provide clues as to how the main components of the video should appear, so that the consistency of visual storytelling is preserved. As such, a scheme to condition the video generation on image not by directly modifying it, but using it as a reference to adjust the visual appearances will be necessary to enable a recurrent video generation, which is more desirable than current scheme of concatenating independently generated scenes.

Script-to-video generation: As discussed in Section 4.2, script is a familiar format for humans to perceive stories. As shown in Section 5.2, nonetheless, generating videos from dialogue is as of now far from delivering a sensible and coherent visuals, which implies a severe asymmetry in terms of the types of text in training corpus. In our attempts, the model frequently fails with who is present in the scene, when given a script as input prompt, as shown in Figure 7. This implies that the current model does not understand how the script format is structured, particularly the role of speaker identification token. Once dialogue-to-video generation is established, a much more powerful storytelling will be possible, as it can be more easily incorporated with audible modality using text-to-speech genera-

Table 5. Examples of human-written summary. The first two summaries are based on "The Lost Melody", and the last two are based on the birthday episode shown in Figure 4.

Model	Human-written summary
	"A little girl is smiling in front of a house in forest. The girl is sitting next to her grandma. The girl is
Gen.(story)	playing a violin in a room. Now, the girl goes outside and starts playing violin again. The girl is lost in
	the forest and people are searching for her. The girl manages to come back home safely. The girl plays
	violin again with joy of returning home."
	"A young girl and grandma are enjoying their time next to fireplace. The girl starts playing a violin. The
Gen.(script)	girl goes out to forest. In the forest, she is fascinated by nature around her. Back at home, the girl and
	her mother start growing flowers. Now she plays violin with a different meaning of loving nature."
Gen.(caption)	"A boy is happy with his birthday cake. The boy opens the door. The boy, with his eyes closed, is
Gen.(caption)	praying in front of the cake with candles lit."
Real	"A happy boy is jumping in front of his birthday cake. The boy opens the door to welcome his friends.
	The boy blows the candles and the friends celebrate."

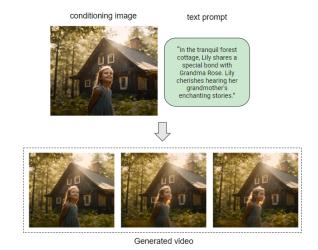


Figure 6. A failure case of video generation conditioned on both image and text inputs.



Figure 7. Generating videos from script format frequently results in complete failure of understanding the scene.

tion, which turned out to be helpful in conveying a narrative in our experiments. Lip-syncing the characters to given text is also an active research topic [21], and can be more easily integrated with videos if dialogue is given.

Global story representation: Our premise so far has been a concatenation of separately generated videos. Ideally, however, a scheme for a global understanding of structure of the story will be necessary. In music generation, for example, a hierarchical architecture is frequently employed

[37] in order to learn the global structure of a musical piece, consisting of ascendance, descent, and repetition of theme, which are also observed in stories. Näively porting the same approach to video domain, however, will require prohibitively large amount of computational resources, and it may thus be necessary to come up with a resource-efficient approach that can compactly encode the structure of a story. One possibility is to represent the structure of a story with graphs, as have been examined by previous works [28, 49]. Incorporating such global representation into video generation can potentially enhance the generation of more coherent videos.

**Evaluation metrics**: While we proposed and conducted a number of evaluation protocols in this paper, they still relied heavily on manual efforts by human. Fully automatic evaluation metrics, such as T2Vid2T, also suffered from still very low similarity between generated captions and ground truth stories, and it also does not account for visual consistency of the character. Such issues may be mitigated by building a more story-oriented video captioning model, which in turn requires a construction of large-scale datasets with videos and corresponding stories. In any case, it is inarguable that more reliable and standardized evaluation metrics for storytelling aspects of videos are indispensable for the field to make substantial advances.

#### 7. Conclusion

We examined current text-to-video generations from a novel storytelling perspective, by generating videos from short stories and scripts, as well as existing captions for video storytelling. We performed an extensive evaluation of generated videos by proposing and conducting a wide array of evaluation protocols. Finally, we made a number of empirical observations and suggestions for potential future directions that we believe would help enhance text-to-video generation for storytelling.

#### References

- [1] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. Spice: Semantic propositional image caption evaluation. *ArXiv*, abs/1607.08822, 2016. 5
- [2] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? ArXiv, abs/2102.05095, 2021. 5
- [3] Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, Clarence Ng, Ricky Wang, and Aditya Ramesh. Video generation models as world simulators. 2024. 2
- [4] João Carreira, Eric Noland, Andras Banki-Horvath, Chloe Hillier, and Andrew Zisserman. A short note about kinetics-600. ArXiv, abs/1808.01340, 2018. 5
- [5] João Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 4724–4733, 2017. 5
- [6] Hong Chen, Duc Vo, Hiroya Takamura, Yusuke Miyao, and Hideki Nakayama. StoryER: Automatic story evaluation via ranking, rating and reasoning. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 1739–1753, Abu Dhabi, United Arab Emirates, Dec. 2022. Association for Computational Linguistics.
- [7] Aidan Clark, Jeff Donahue, and Karen Simonyan. Adversarial video generation on complex datasets. *arXiv: Computer Vision and Pattern Recognition*, 2019. 5
- [8] Michael J. Denkowski and Alon Lavie. Meteor universal: Language specific translation evaluation for any target language. In WMT@ACL, 2014. 5
- [9] Ming Ding, Wendi Zheng, Wenyi Hong, and Jie Tang. Cogview2: Faster and better text-to-image generation via hierarchical transformers. ArXiv, abs/2204.14217, 2022. 2
- [10] Patrick Esser, Johnathan Chiu, Parmida Atighehchian, Jonathan Granskog, and Anastasis Germanidis. Structure and content-guided video synthesis with diffusion models. 2023 IEEE/CVF International Conference on Computer Vision (ICCV), pages 7312–7322, 2023. 4
- [11] Zach Evans, CJ Carr, Josiah Taylor, Scott H. Hawley, and Jordi Pons. Fast timing-conditioned latent audio diffusion. ArXiv, abs/2402.04825, 2024. 4
- [12] Jian Guan and Minlie Huang. UNION: An Unreferenced Metric for Evaluating Open-ended Story Generation. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 9157–9166, Online, Nov. 2020. Association for Computational Linguistics. 2
- [13] Jian Guan, Zhexin Zhang, Zhuoer Feng, Zitao Liu, Wenbiao Ding, Xiaoxi Mao, Changjie Fan, and Minlie Huang. Open-MEVA: A benchmark for evaluating open-ended story generation metrics. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, Proceedings of the 59th Annual Meeting of the Association for Computational Linguis

- tics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 6394–6407, Online, Aug. 2021. Association for Computational Linguistics. 2
- [14] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey A. Gritsenko, Diederik P. Kingma, Ben Poole, Mohammad Norouzi, David J. Fleet, and Tim Salimans. Imagen video: High definition video generation with diffusion models. *ArXiv*, abs/2210.02303, 2022. 2
- [15] Jonathan Ho, Ajay Jain, and P. Abbeel. Denoising diffusion probabilistic models. ArXiv, abs/2006.11239, 2020. 2
- [16] Wenyi Hong, Ming Ding, Wendi Zheng, Xinghan Liu, and Jie Tang. Cogvideo: Large-scale pretraining for text-to-video generation via transformers. ArXiv, abs/2205.15868, 2022.
- [17] Ting-Yao Hsu, Chieh-Yang Huang, Yen-Chia Hsu, and Ting-Hao Huang. Visual story post-editing. In Anna Korhonen, David Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6581–6586, Florence, Italy, July 2019. Association for Computational Linguistics. 3
- [18] Junjie Hu, Yu Cheng, Zhe Gan, Jingjing Liu, Jianfeng Gao, and Graham Neubig. What makes a good story? designing composite rewards for visual storytelling. In AAAI Conference on Artificial Intelligence, 2019. 3, 5
- [19] Ting-Hao 'Kenneth' Huang, Francis Ferraro, N. Mostafazadeh, Ishan Misra, Aishwarya Agrawal, Jacob Devlin, Ross B. Girshick, Xiaodong He, Pushmeet Kohli, Dhruv Batra, C. Lawrence Zitnick, Devi Parikh, Lucy Vanderwende, Michel Galley, and Margaret Mitchell. Visual storytelling. In North American Chapter of the Association for Computational Linguistics, 2016. 2, 5
- [20] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. Large-scale video classification with convolutional neural networks. 2014 IEEE Conference on Computer Vision and Pattern Recognition, pages 1725–1732, 2014. 5
- [21] Rithesh Kumar, Jose M. R. Sotelo, Kundan Kumar, Alexandre de Brébisson, and Yoshua Bengio. Obamanet: Photorealistic lip-sync from text. *ArXiv*, abs/1801.01442, 2017.
- [22] Junnan Li, Yongkang Wong, Qi Zhao, and M. Kankanhalli. Video storytelling: Textual summaries for events. *IEEE Transactions on Multimedia*, 22:554–565, 2018. 5
- [23] Y. Li, Z. Gan, Y. Shen, J. Liu, Y. Cheng, Y. Wu, L. Carin, D. Carlson, and J. Gao. Storygan: A sequential conditional gan for story visualization. In 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 6322–6331, Los Alamitos, CA, USA, jun 2019. IEEE Computer Society. 2
- [24] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In Annual Meeting of the Association for Computational Linguistics, 2004. 5
- [25] Han Lin, Abhaysinh Zala, Jaemin Cho, and Mohit Bansal. Videodirectorgpt: Consistent multi-scene video generation via llm-guided planning. ArXiv, abs/2309.15091, 2023. 2
- [26] Danyang Liu, Mirella Lapata, and Frank Keller. Visual storytelling with question-answer plans. In Houda Bouamor,

- Juan Pino, and Kalika Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5800–5813, Singapore, Dec. 2023. Association for Computational Linguistics. 3
- [27] Fuchen Long, Zhaofan Qiu, Ting Yao, and Tao Mei. Video-drafter: Content-consistent multi-scene video generation with llm. ArXiv, abs/2401.01256, 2024.
- [28] Stephanie M. Lukin, Kevin K. Bowden, Casey Barackman, and Marilyn A. Walker. Personabank: A corpus of personal narratives and their story intention graphs. ArXiv, abs/1708.09082, 2016. 8
- [29] Adyasha Maharana, Darryl Hannan, and Mohit Bansal. Storydall-e: Adapting pretrained text-to-image transformers for story continuation. *ArXiv*, abs/2209.06192, 2022. 7
- [30] OpenAI. Openai: Introducing chatgpt. https://openai.com/blog/chatgpt, 2022. 3
- [31] OpenAI. Dall·e 3. https://openai.com/dall-e-3, 2023. 2
- [32] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Annual Meeting of the Association for Com*putational Linguistics, 2002. 5
- [33] Zengyi Qin, Wenliang Zhao, Xumin Yu, and Xin Sun. Openvoice: Versatile instant voice cloning. ArXiv, abs/2312.01479, 2023. 4
- [34] 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. In *ICML*, 2021.
- [35] Colin Raffel, Noam M. Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. J. Mach. Learn. Res., 21:140:1–140:67, 2019.
- [36] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *ArXiv*, abs/2204.06125, 2022. 2
- [37] Adam Roberts, Jesse Engel, Colin Raffel, Curtis Hawthorne, and Douglas Eck. A hierarchical latent vector model for learning long-term structure in music. ArXiv, abs/1803.05428, 2018. 8
- [38] Masaki Saito, Eiichi Matsumoto, and Shunta Saito. Temporal generative adversarial nets with singular value clipping. 2017 IEEE International Conference on Computer Vision (ICCV), pages 2849–2858, 2016. 5
- [39] Tim Salimans, Ian J. Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. ArXiv, abs/1606.03498, 2016. 5
- [40] Uriel Singer, Adam Polyak, Thomas Hayes, Xiaoyue Yin, Jie An, Songyang Zhang, Qiyuan 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. ArXiv, abs/2209.14792, 2022. 2
- [41] Khurram Soomro, Amir Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. ArXiv, abs/1212.0402, 2012. 5

- [42] Du Tran, Lubomir D. Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. 2015 IEEE International Conference on Computer Vision (ICCV), pages 4489–4497, 2014. 5
- [43] Thomas Unterthiner, Sjoerd van Steenkiste, Karol Kurach, Raphaël Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric & challenges. *ArXiv*, abs/1812.01717, 2018. 5
- [44] Aäron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning. ArXiv, abs/1711.00937, 2017. 2
- [45] Ashish Vaswani, Noam M. Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017.
- [46] Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 4566–4575, 2014. 5
- [47] Eileen Wang, Caren Han, and Josiah Poon. RoViST: Learning robust metrics for visual storytelling. In Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz, editors, Findings of the Association for Computational Linguistics: NAACL 2022, pages 2691–2702, Seattle, United States, July 2022. Association for Computational Linguistics. 3
- [48] Xin Wang, Wenhu Chen, Yuan-Fang Wang, and William Yang Wang. No metrics are perfect: Adversarial reward learning for visual storytelling. In Iryna Gurevych and Yusuke Miyao, editors, Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 899–909, Melbourne, Australia, July 2018. Association for Computational Linguistics. 3
- [49] Liyan Xu, Jiangnan Li, Mo Yu, and Jie Zhou. Graph representation of narrative context: Coherence dependency via retrospective questions. 2024. 8

# A. Appendix

#### A.1. Short Story Generation

#### **Prompt:**

Write a plot for a 5-minute animation. It should involve basic elements of a story, including characters, setting, plot, conflict, and theme. The overall mood should be warm and heartfelt, but it may be good to add an element of eeriness.

#### **Output:**

Title: "The Lost Melody"

Setting: A cozy, rustic cottage in a peaceful forest. Characters:

Lily: A young girl with a bright smile, staying at the cottage with her grandmother. Grandma Rose: A kind and wise elderly woman who lives in the cottage. The Enchanted Violin: An old, mysterious violin found in the attic. Plot:

The story begins with Lily visiting her Grandma Rose in the tranquil cottage deep in the forest. They share a special bond, and Lily loves hearing her grandmother's stories.

One day, while exploring the attic, Lily stumbles upon an old, dusty violin. Grandma Rose explains that it's an enchanted violin, but it has been silent for years because it's missing a magical melody.

Intrigued by the idea of an enchanted violin, Lily is determined to find the lost melody. She embarks on a quest through the forest, seeking magical elements to create the melody. Along the way, she encounters talking animals who offer clues, and the forest itself seems to come to life.

As Lily's journey unfolds, an eerie, mystical atmosphere envelopes the forest. Strange and enchanting occurrences, like glowing fireflies and ethereal music, add an element of eeriness to the story.

Finally, with the help of the forest creatures, Lily finds all the ingredients needed for the magical melody and returns to the cottage.

She plays the melody on the enchanted violin. It responds with beautiful, heartwarming music that fills the forest, causing flowers to bloom and animals to dance.

The message of the story becomes clear: The magic is in the connection between generations and the beauty of passing down traditions and stories.

Theme: The central theme of "The Lost Melody" is the power of tradition, the importance of family bonds, and the enchanting wonder that can be found in even the most unexpected places. It combines warmth and heartfelt moments with an element of mystery and magic, creating an unforgettable and heartwarming tale.

### A.2. Prompt Sequence Generation

#### **Prompt:**

I want to generate a video for the generated short story above using AI text-to-video generation tools. So I need to divide them into a sequence of texts, each of which is to be the prompt for the generation. For the prompt to successfully generate coherent video, each prompt will have to contain the character and setting, since each prompt is independent of others from the tool's perspective. Please make the prompt sequences.

### **Output:**

Introduction to Setting and Character:

"Meet Lily, a young girl with a bright smile, staying at a cozy, rustic cottage nestled in a peaceful forest. Lily is visiting her grandmother, Grandma Rose."

Special Bond and Grandma's Stories: "In the tranquil forest cottage, Lily shares a special bond with Grandma Rose. Lily cherishes hearing her grandmother's enchanting stories."

Discovery of the Enchanted Violin: "While exploring the attic one day, Lily discovers an old, dusty violin in the cozy

cottage. Grandma Rose reveals its mystery – it's an enchanted violin, silent for years without its magical melody."

Quest for the Lost Melody Begins: "Intrigued by the enchanted violin, Lily is determined to find the lost melody. She embarks on a quest through the mystical forest surrounding the cozy cottage, seeking magical elements to create the melody."

Encounters in the Enchanted Forest: "As Lily explores the enchanted forest, she encounters talking animals who offer clues. The forest itself seems to come to life around the cozy cottage. Strange occurrences, like glowing fireflies and ethereal music, add an eerie yet enchanting atmosphere."

Culmination of Lily's Journey: "During Lily's journey, an eerie, mystical atmosphere envelops the forest around the cozy cottage. With the help of the forest creatures, Lily gathers all the magical ingredients and returns to the cottage."

Magical Melody Unveiled: "Back at the cozy cottage, Lily plays the magical melody on the enchanted violin. The violin responds with beautiful, heartwarming music that fills the forest. Flowers bloom, and animals dance in response."

The Message of the Story: "In this enchanting tale, discover the magic in the connection between generations and the beauty of passing down traditions and stories at the cozy, rustic cottage in the peaceful forest." Now, each prompt contains both the character (Lily, Grandma Rose) and the setting (cozy cottage in the peaceful forest). Adjustments can still be made based on the specific requirements of the text-to-video generation tool.

### A.3. Script Generation

#### **Prompt:**

Convert the story above into a script with dialogue to be spoken by characters, with non-dialogue part for indicating setting, transition, etc. Write the script as a prompt sequence for AI text-to-video generation tools, while including characters in each prompt as you did before.

#### **Output:**

INT. COZY COTTAGE - DAY

Setting: A cozy, rustic cottage in a peaceful forest. Characters: Lily (young girl), Grandma Rose (elderly woman) INT. COZY COTTAGE - LIVING ROOM - DAY

Lily and Grandma Rose sit by the fireplace. Dialogue: Lily: "Grandma, your stories always make the cottage feel even cozier." Grandma Rose: "Well, my dear, stories have a way of adding warmth to our surroundings." INT. COZY COTTAGE - ATTIC - DAY

Lily explores the attic, discovering an old, dusty violin. Dialogue: Lily: "What's this, Grandma?" Grandma Rose: "Ah, that, Lily, is no ordinary violin. It's enchanted, but it has been silent for years." EXT. COZY COTTAGE - DAY

Lily sets out on a quest through the mystical forest. INT/EXT. ENCHANTED FOREST - VARIOUS LOCATIONS - DAY

Lily encounters talking animals, gathers magical elements, and experiences the mystical atmosphere of the forest. Dialogue: Lily (to a talking bird): "Have you heard of the lost melody? I need your help." Talking Bird (chirping): "Follow the fireflies. They hold the secret." EXT. COZY COTTAGE - DAY

Lily returns to the cottage, the enchanted forest echoing with mysterious sounds. INT. COZY COTTAGE - LIVING ROOM - DAY

Lily, with gathered ingredients, stands before Grandma Rose. Dialogue: Lily: "Grandma, I'm going to play the lost melody on the enchanted violin!" Grandma Rose: "Oh, Lily, that melody has been silent for far too long. Let the forest hear its song again." EXT. COZY COTTAGE - DAY

Lily plays the magical melody. The forest responds with blooming flowers and dancing animals. INT. COZY COTTAGE - LIVING ROOM - DAY

Grandma Rose watches with joy. Dialogue: Grandma Rose (whispering): "The magic is in the connection between generations, Lily. You've brought the melody back to life." EXT. COZY COTTAGE - DAY

The forest, now alive with enchantment, surrounds the cozy cottage.