Measuring The Consistency of Image Visualisations

Maksymilian Ćwirzeń

MSc Artificial Intelligence

The University of Bath

2024

This dissertation may be made available for consultation within the University Library and may be photocopied or lent to other libraries for the purposes of consultation.

Measuring The Consistency of Image Visualisations

Submitted by: Maksymilian Ćwirzeń

Copyright

Attention is drawn to the fact that copyright of this dissertation rests with its author. The Intellectual Property Rights of the products produced as part of the project belong to the author unless otherwise specified below, in accordance with the University of Bath's policy on intellectual property (see https://www.bath.ac.uk/publications/university-ordinances/attachments/Ordinances_1_October_2020.pdf).

This copy of the dissertation has been supplied on condition that anyone who consults it is understood to recognise that its copyright rests with its author and that no quotation from the dissertation and no information derived from it may be published without the prior written consent of the author.

Declaration

This dissertation is submitted to the University of Bath in accordance with the requirements of the degree of MSc Artificial Intelligence in the Department of Computer Science. No portion of the work in this dissertation has been submitted in support of an application for any other degree or qualification of this or any other university or institution of learning. Except where specifically acknowledged, it is the work of the author.

Abstract

Amidst the rise of robust text-to-image models in the field of Generative Artificial Intelligence, one subfield has received relatively little attention from the public and the science community as its development hasn't been able to keep up with the rest of the field. The field of **Consistent Image Visualization** is an extension of a wider Image Visualisation field with an added objective of generating a series of images whereby their shared features are represented consistently. For a long time, one of the major problems of this field has been a lack of an objective evaluation metric for measuring the consistency between generated images. Due to this, researchers in this area were forced to resort to user studies when analysing their results, while others used substitute evaluation metrics such as FID, Char-F1 and Character Accuracy, leading to significant deficiencies in gathered quantitative data. This thesis identifies and attempts to resolve this problem by introducing a new metric called **F2SD** (Fréchet Subject Similarity Distance), aiming to quantify the relations between source and target frames generated by text-to-image consistency models as well as eliminate the need for costly user studies.

Contents

Acknowledgements

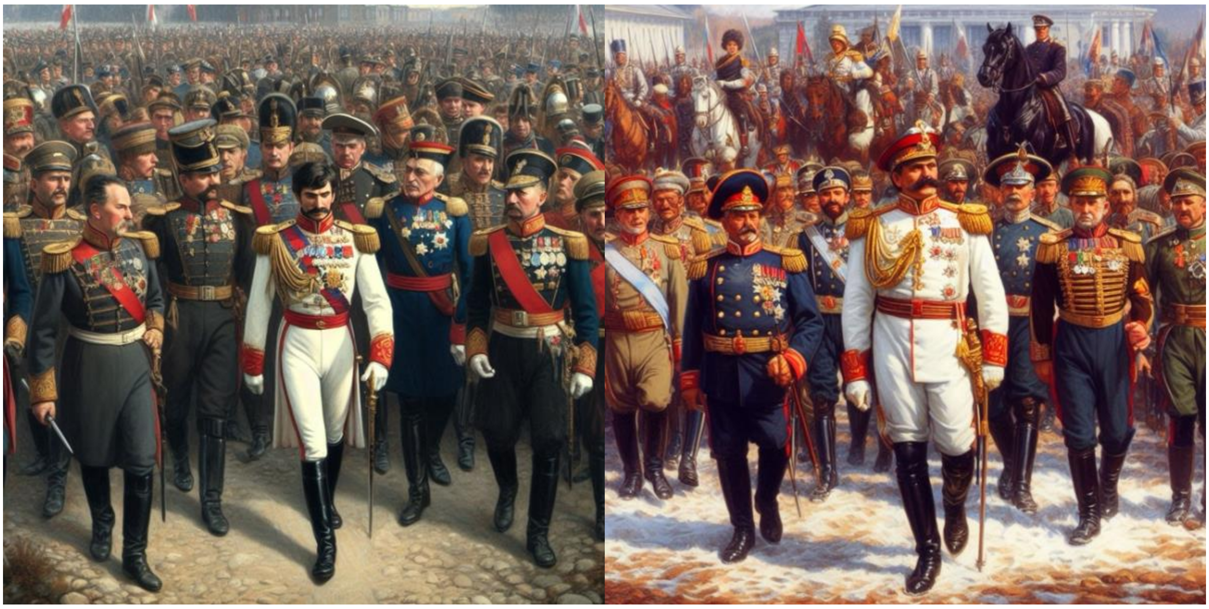
It is my greatest pleasure to thank Dr Raghubir Singh for supervising this project and for all the support and guidance that he afforded me. I would like to further thank all lecturers and staff at the University of Bath that I had the privilege to interact with, whose hard work created the environment in which I was able to expand my horizons and complete this course. Naturally, my thanks extend to all my friends and family who supported me along this journey – I am extremely grateful to you all. Lastly, I would like to thank the Internet without which attending this course wouldn’t be possible.

Chapter 1: Introduction

Background

Stories have always been a fundamental part of the human experience. They accompany us since our very beginnings. Read by parents, they are a key part of our brains' development. Read to children, stories aim to instil fundamental values like empathy, friendship and assertiveness. For adults, they intent to convey ideas, entertain and encourage, and once we get old – make us reflect, bring back in time, feel younger. Storytelling is universal to the human experience and although it is hard to prove, historians suggest that this practice have developed shortly after the development of language itself (National Geographic Society, 2023). For thousands of years, storytelling was limited to oral traditions passed on from generation to generation, which gave birth to mythologies, legends and folk tales. With development of written text, people began putting their stories onto stone and clay tablets, tree bark, eventually moving onto papyrus, parchment and paper. The invention and spread of the printing press enabled global literacy and the global reach of written stories skyrocketed. Thanks to the technological advancements in the past century, today's storytelling has many new visual dimensions that are not limited to text or static illustrations. We have movies, audiobooks, comics, documentaries, and many new interactive forms like video or role-playing games. Seeing how the storytelling evolved over the thousands of years is a testament to how important new technologies are in shaping the way we see the world.

Following this trend, the arrival of Artificial Intelligence opened up a new areas for exploration and experimentation. As early as in 1977, we saw TALE-SPIN – the first reported narrative generation system (Meehan, 1977). Then came Ani (Kahn, 1979) and UNIVERSE (Lebowitz, 1985). These systems were very basic and focused on fitting pre-existing plot fragments from simple interactions between random characters. Current work in this area of AI has been mostly focused on the content generation for children as it assumes a form of short storybook generation (Ansag & Gonzalez, 2021). Some examples include the use of pre-determined plot patterns such as the CAMPFIRE Storytelling System (Hollister & Gonzalez, 2018), a stochastic model such as AESOP (Wade, et al., 2017), a combination of a stochastic model and a graphical database aiming to avoid irrelevant story actions and increase its variety (Bottoni, et al., 2020) or combined text and visualisation generation (storybookai, 2023). Today, in the era of generative AI, large language models enable effortless story generation and visualisation that is available to the public. It takes less than a few minutes to illustrate a scene from a book using ChatGPT (OpenAI, 2023) as shown in the Figures 1 & 2.



Figures 1 & 2: Page 157 of War and Peace summarized and visualised by GPT-4 (OpenAI, 2023). "Kutuzov walking through the ranks, officers and soldiers around him, with Prince Bolkonski and Nesvitski beside him, and a hussar officer mimicking the regimental commander's movements." (Tolstoy, 1869)

Motivation

OpenAI's GPT, DALL-E and the Midjourney models have recently became by far the most popular tools for story generation. As described by Hariffadzillah et al., they can play a pivotal role in not only producing children literature, but any literature in general (Hariffadzillah, et al., 2023) thanks to their incredible capabilities. Models designed specifically for story generation can't achieve such high level of abstraction as LLM's do thanks to their universality. It is likely that we will see a further increase in machine capabilities in this area as a by-product of the currently ongoing inter-corporational race for the best LLM.

Another form of literature that is often being put to a practical use are storyboards. They are being used in marketing and advertising industries during the writing process to help pre-visualise scene sequences. Sadly, there isn't much in terms of relevant and reliable written literature for storyboard generation. Nevertheless, there many applications available on the web that advertise themselves as storyboard generators. Most of them offer the same core set of features such as robust storyboard generation from a concise scene description prompt. They can offer various different art styles like krock.io (krock.io, 2023), offer a simplistic and friendly UI like boords (boords, 2023) or offer faster iteration through different versions of the scene like StoryBoardHero (StoryboardHero, 2023). In their core, they are all very similar products that are most likely using either DALL-E or

MidJourney models in the backend to generate results (although none of them state this explicitly).

Despite that, the above applications still remain niche and don't get much attention from the scientific community, and the general public. One of the answers as to why this might be the case is the matter of **consistency**. Currently available models aren't capable of "remembering" our story subjects very well. Not to their discredit – they are simply not made to do so. This takes us to the main topic of this thesis – the field of **Consistent Image Visualization** (or **Story Visualisation**), of which aim is to explore various ways to enable modern image generation models to "remember" various relevant aspects of previous images (called frames) and successfully replicate them in different contexts.

Problem Statement and Research Question

Initially, this project's aim was to investigate different types of visualisations, based upon different types of abstract text and create a proof of concept for automation of certain creative processes that take place in the filmmaking industry. Upon further reading, this has since evolved into a wide survey aiming to connect different tropes in the fields of Consistent Image Visualisation and Story Visualisation that were identified to be fully or partially disconnected from one another.

This thesis is looking to answer the following question:

- "How can the different tropes in the field of Consistent Image Visualisation be brought together?"

In this work, we introduce **Fréchet Subject Similarity Distance (F2SD)** – the first **subject consistency evaluation metric** that can aid the researchers who possess the available resources to advance the field of story visualization and elevate storytelling to a new high in the coming years. We propose it as a solution to the research question above thanks to it being a first known metric measuring distance between two generated images, regardless of their relations to ground truth.

Chapter 2: Literature Review and Technology Survey

Prerequisites

Attention Is All You Need

Natural Language Processing (NLP) is one of the oldest fields of Artificial Intelligence. If an underlying goal of the AI research is to reach Artificial General Intelligence – it is going to be hard to imagine such model without a very advanced NLP component. After all, language is a way humans communicate and convey their ideas and emotions. Therefore, NLP is an interdisciplinary field that investigates structures and features of the human language, the way it's being used in practice and attempts to create language models that could emulate its real-world use (DeepLearning.AI, 2023).

Most recently, we've seen the rise of neural-network based approaches in form of Large Language Models (LLM) such as various OpenAI's GPT models (OpenAI, 2023) or Meta's open-source Llama family of models (Meta, 2023). Modern day LLM's use the transformer architecture. This approach has taken over the field after years of domination of statistical language models based on word embeddings followed by Recurrent Neural Network (RNN) based models and various other Long Short-Term Memory (LSTM) based models. Transformers in their simplest forms are attention-based sequential encoder-decoder models whose potential was fully realised by Vaswani et al. – a group of researchers working for Google in 2017.

The biggest problem at a time was designing a language model architecture that would reduce the training costs and improve the efficiency. The attention mechanism was first introduced by Bahdanau et al. to improve the performance of models used for language translation (Bahdanau, et al., 2014). They initially proposed it to be a part of a RNN model. However, Vaswani et al. were the first to completely disregard recurrence and convolution and attempted to build a network composed entirely of attention modules. Their model vastly outperformed current state-of-the-art recurrence-based models by increasing parallelisation and therefore significantly decreasing the training time and runtime (Vaswani, et al., 2017). Their architecture was also much more scalable which enabled other researchers to train much better language models faster. This work has since become a landmark in the field of AI enabling the recent "AI-boom".

The attention function Vaswani et al. use in their work is called Scaled Dot-Product Attention which they described with the following equation:

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

Where Q and K are vectors of queries and keys of size d_k respectively and V is a vector of values. The researchers described the process as 'computing the output as a weighted sum of the values, where the weight assigned to each value is computed by a compatibility function of the query with the corresponding key'. In summary, Vaswani et al.'s solution was to develop a network with a series of dynamic weights dependant on another set of inputs. This model is much better suited to capturing relationships within input sequences as it allows for a much better context encoding.

As it turned out, Natural Language Processing was not the only field of AI that would benefit from this feature.

Latent Diffusion

Image Generation is a field of Artificial Intelligence that has recently attracted a lot of attention thanks to extremely impressive results achieved by models like DALL·E (OpenAI, 2023) and Midjourney (MidJourney, 2022). These models belong to a family of Latent Diffusion models. The diffusion process in the context of AI has been introduced by Sohl-Dickstein et al. in 2015. The method they came up with was to iteratively destroy structures in data distribution by adding noise and train a network to learn how to reverse this process. This method results in a model which is highly flexible and tractable (Sohl-Dickstein, et al., 2015). The greatest downside of this method was that it was computationally inefficient due to a very expensive evaluation function. The models were operating in pixel spaces which required hundreds of GPU-days to complete the training. This is where latent diffusion comes in. The concept was first introduced by Rombach et al., a group of researchers from the University of Münich. They encoded the pixel space using a pretrained autoencoder network and achieved near-optimal results with a much smaller computational cost. The encoded latent space is perceptually equivalent to the pixel space but offers significantly reduced computational complexity (Rombach, et al., 2021). Its loss function can be mathematically expressed as:

$$L_{LDM} \coloneqq \mathbb{E}_{\mathcal{E}(x),\epsilon\sim\mathcal{N}(0,1),t}[||\epsilon - \epsilon_\theta(z_t, t)||_2^2],$$

For image x, encoder $\mathcal{E}$, encoded data z, distribution $\mathcal{N}$, step t, noise ϵ and the network ϵ_θ.

Thanks to this revelation, diffusion models have taken over the field, breaking a long dominance of Generative Adversarial Networks (GANs) by offering greater control and stability during training (Dhariwal & Nichol, 2021), and by reducing model complexity. On top of that, as described in a recent article by Ling et al., diffusion models are showing great potential in many other domains and applications ranging from computer vision, temporal data modelling, multi-modal modelling, computational chemistry, medical imaging and Natural Language Processing (NLP) (Ling, et al., 2023).

What constitutes a good prompt?

Coming up with the right prompt is an essential part towards achieving desired visualisations. A lot of previous related work that has been done in this matter leaves the prompt generation to the user. This means that the results rely heavily on the user's knowledge and intuition. There are many factors that make the prompt good. Here are some core attributes:

1. **Specificity**. Provide enough information to describe the subject, setting, artistic style, and other pertinent aspects clearly and accurately.
2. **Conciseness**. Keep the prompt focused and free of extraneous content that might distract the model or introduce unwanted variations.
3. **Relevance**. Ensure that included keywords relate specifically to the concept you intend to convey.
4. **Unambiguity**. Choose terminology that leaves minimal room for misinterpretation.

This is often not enough, and many diffusion applications are also using negative prompts to eliminate unwanted results. Researchers have been looking for ways to further enhance prompt understanding. For example, Lian et al. take advantage of an LLM module to generate a series of layouts and 'sub-prompts' as shown on the Figure 3 below (Lian, et al., 2023).

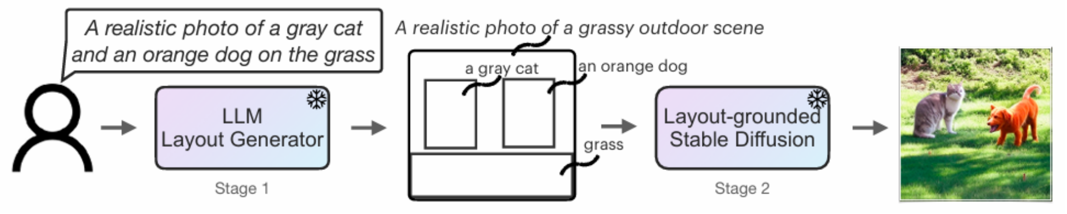

A realistic photo of a gray cat and an orange dog on the grass
LLM Layout Generator
Stage 1
A realistic photo of a grassy outdoor scene
a gray cat
an orange dog
grass
Layout-grounded Stable Diffusion
Stage 2

Figure 3: LLM-grounded Diffusion pipeline (Lian, et al., 2023).

Story Visualisation

The arrival of robust image generation models paved the way for all sorts of visual story generation products - most notably storybookai (storybookai, 2023). The researchers have quickly come up against a serious problem that the image generation models were not suited to deal with – **consistency**. In the context of this field, it means generating a series visualisations such that the visualised subjects are consistently projected across all frames.

Consistency is all you need

The Naming Conundrum

Before we begin analysing the subject matter it's worth stopping for a second to analyse a grossly overlooked problem that the research field has been struggling with and that seems to have gone largely unnoticed or unacknowledged. The problem begins with the realisation that there is no unified understanding on what the field should be called. This sounds trivial but it poses a serious research problem.

Example:

- Researcher $\boldsymbol{A}$ is interested in using AI to generate story visualizations. He's struggling to make sure that the subjects of the story are consistently generated (among all other things). He goes to Google Scholar and types "Consistent Story Visualisation" and finds a set of works $\boldsymbol{X}$.
- Researcher $\boldsymbol{B}$ is interested in generating images of a certain subject. He's struggling to make his model generate said subjects consistently across multiple frames. He types "Consistent Image Generation" into Google Scholar and finds a set of works $\boldsymbol{Y}$.

Both sets $\boldsymbol{X}$ and $\boldsymbol{Y}$, define the same problem and present their own solutions; however, they reach those solutions largely independently to each other due to their slightly different problem definitions. Furthermore, an attentive researcher will notice a wide range of phrases denoting the same concept – story continuation, coherent visualisation, consistent generation, etc.

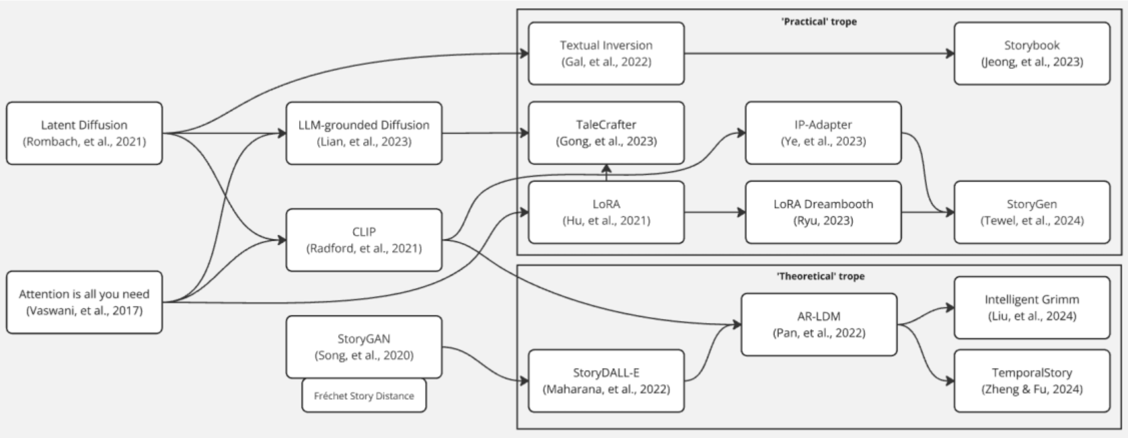

'Practical' trope
Textual Inversion (Gal, et al., 2022)
Storybook (Jeong, et al., 2023)
Latent Diffusion (Rombach, et al., 2021)
LLM-grounded Diffusion (Lian, et al., 2023)
TaleCrafter (Gong, et al., 2023)
IP-Adapter (Ye, et al., 2023)
LoRA (Hu, et al., 2021)
LoRA Dreambooth (Ryu, 2023)
StoryGen (Tewel, et al., 2024)
CLIP (Radford, et al., 2021)
'Theoretical' trope
Attention is all you need (Vaswani, et al., 2017)
AR-LDM (Pan, et al., 2022)
Intelligent Grimm (Liu, et al., 2024)
StoryGAN (Song, et al., 2020)
Fréchet Story Distance
StoryDALL-E (Maharana, et al., 2022)
TemporalStory (Zheng & Fu, 2024)

Figure 4: Survey Organisation Diagram depicting approximate relations between relevant works.

Most importantly, the works in the field are divided along the line of how they achieve desired results. This divide can be plainly seen on Figure 4 representing the Survey Organisation Diagram. On one side (bottom rectangle), we have teams working with PororoSV and FlintstonesSV datasets, achieving impressive yet very theoretical breakthroughs. Those models are trained against and compare to an objective ground truth given by the dataset. The other side (top rectangle) focuses on consistency in a more practical sense and comes up with methods that can be implemented in conjunction with the existing diffusion models. As opposed to the aforementioned models, these operate with no reference to an objective ground truth. Due to this, they are far less specialised, more flexible and applicable. Not to say that one is better than the other, but it needs to be acknowledged that there is a big disconnect between the two tropes and given the fact that both are ultimately working towards the same goal – we identify the need of bringing them closer by compiling a comprehensive survey of the field, introducing more quantitative results along the way and implementing an evaluation metric that would be applicable to both schools of thought. In order to make this thesis easier to read, we will refer to the two approaches as ‘theoretical’ and ‘practical’ tropes from now on.

“Theoretical” trope

One of the first network-based non-GAN solutions for consistent visualisation was introduced in 2022 by Maharana et al. Dubbed **STORYDALL-E**, its creators took existing GAN-based solutions and adapted them onto an up-and-coming Latent Diffusion technique. They took an open-source minDALL-E model (Kuprel, 2022) and modified it such that it retroactively fitted the **source frame** (the initial frame) to all **target frames** (subsequent frames) (Maharana, et al., 2022). The researchers seek to improve the time performance of the model by encoding the input using a self-attention network instead of a recurrent LSTM network. Maharana et al. were among the first to combine Transformer and Diffusion architectures to create what is now called a Diffusion Transformers (DiTs) class of models (termed as such soon after (Peebles & Xie, 2023)). They further enhanced the encoding by initialising sinusoidal positional embeddings which improved the model’s understanding of the positioning of the target frame within the story. Lastly, the researchers avoided costly full-model fine-tuning by attaching a prompt-tuning network and training that one instead. The resulting model achieved a good level of consistency and with the use of Latent Diffusion, it outclassed old GAN models in image quality as shown on Figure 5 and Table 1.

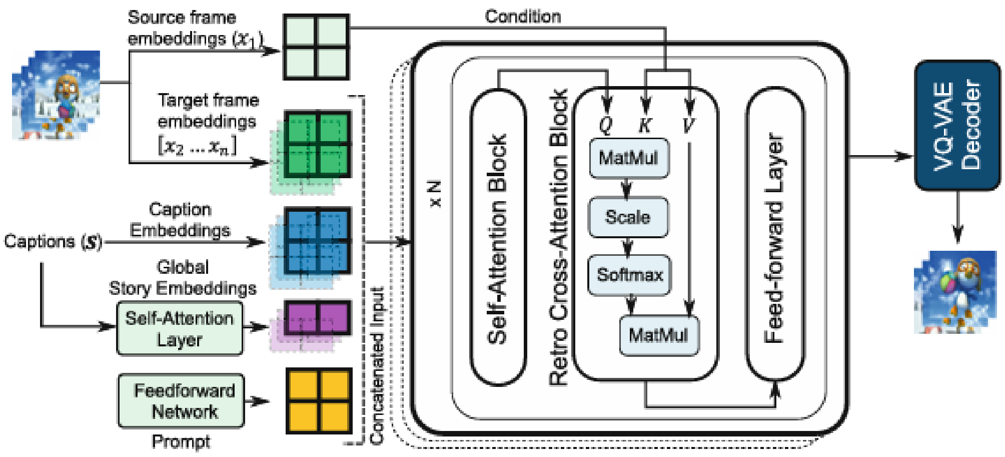


Figure 5: STORYDALL-E model architecture (Maharana, et al., 2022).

STORYDALL-E achieved a great success in this field due to the introduction of transformers which radically boosted image comprehension as they are far better in understanding various abstract and semantically complex concepts. Researchers identified that existing story visualisation solutions were highly dependent on the information provided by the narrative and could only recreate characters already found in the training dataset. The adaptation of attention mechanism into an image generation model allows for a wider range of interpretations and enable greater subject consistency.

Model	**PororoSV**			**FlintstonesSV**		
	FID↓	**Char-F1↑**	**F-Acc↑**	**FID↓**	**Char-F1↑**	**F-Acc↑**
StoryGANc (BERT)	72.98	**43.22**	17.09	91.37	70.45	55.78
StoryGANc (CLIP)	74.63	39.68	16.57	90.29	72.80	**58.39**
StoryDALL-E	**25.90**	36.97	**17.26**	**26.49**	**73.43**	55.19

Table 1: STORYDALL-E performance against StoryGANc (Maharana, et al., 2022).

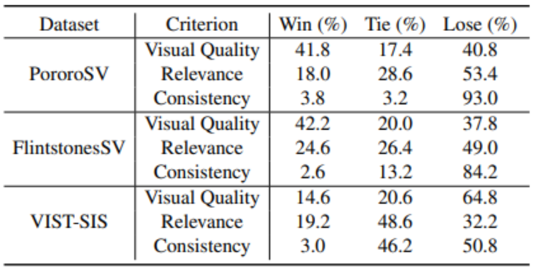
One of the weaknesses that the authors of STORYDALL-E did not identify was that their model was characterised by a low variation in background generation. One can think of this as the transformer architecture working too well. This will turn out to be a common theme among works following the 'theoretical' approach, as many of the subsequent works struggled with the same problem. It was later solved by Tewel et al. in a work which will be described further down in the 'practical' section. They dubbed it more broadly as **information leakage**. It can be explained by a high degree of dependence on the source frame leading to all of its information being considered in generation of target frames. Some information may be desired such as subject's features and attributes, some however may not. These would include things like background, subject's pose, artistic theme, etc. Lastly, Maharana et al. noted that diffusion-generated images were far from being real-world quality. Indeed, what follows was an exercise in improving upon STORYDALL-E's consistency mechanisms, battling its shortcomings and adapting them to newer and better diffusion models.

Another breakthrough happened soon after with the publication of works conducted by Pan et al. They designed an Auto-Regressive Latent Diffusion Model (**AR-LDM**). Their model attempts to break an assumption that each target frame is conditionally independent on previous frames. They achieve that by introducing a novel BLIP network to encode previous caption-frame pairs (Li, et al., 2022) and concatenate with current caption CLIP network encoding (Radford, et al., 2021). This network (so called History-Aware Conditioning Network) is used as conditioning for the diffusion process (Pan, et al., 2022). Furthermore, the researchers seek to alleviate the problem of preserving the consistency of yet unseen characters by extending their model and adding a special embedding token and finetuning parameters on a single story containing such token.

Models	# of Params	FID↓	
		PororoSV	FlintstonesSV
StoryGANc (Maharana, et al., 2022)	-	74.63	90.29
StoryDALL-E (Maharana, et al., 2022)	1.3B	25.90	26.49
MEGA-StoryDALL-E (Maharana, et al., 2022)	2.8B	23.48	23.58
AR-LDM (Pan, et al., 2022)	1.5B	17.40	19.28

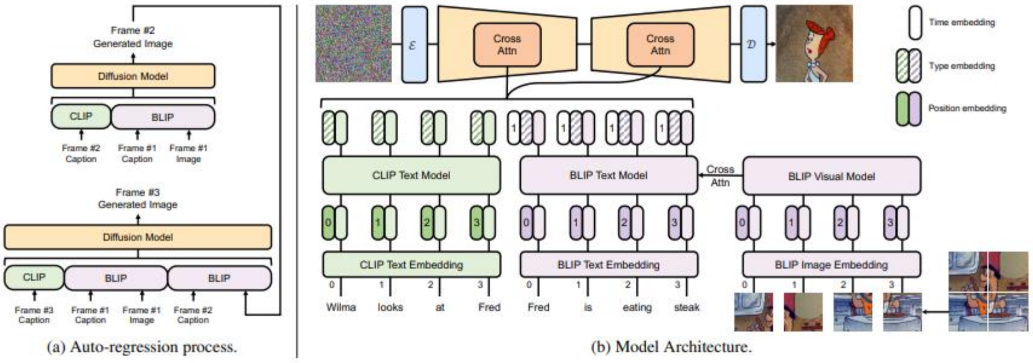
Table 2: AR-LDM image quality (FID) performance against STORYDALL-E (Pan, et al., 2022).

As per Tables 2 & 3, AR-LDM significantly outscores STORYDALL-E in all metrics thanks to the use of a more sophisticated conditioning network. However, the authors note that when comparing consistency against the ground truth, the user studies reveal that the results are still overwhelmingly against the model. Nevertheless, with its publication AR-LDM became the state-of-the-art model for high quality consistent image. Reaching so far unprecedented results through diligent adaptation of most recent technologies like CLIP (Radford, et al., 2021) and BLIP (Li, et al., 2022) encoders, to its advantage.



Dataset	Criterion	Win (%)	Tie (%)	Lose (%)
PororoSV	Visual Quality	41.8	17.4	40.8
	Relevance	18.0	28.6	53.4
	Consistency	3.8	3.2	93.0
FlintstonesSV	Visual Quality	42.2	20.0	37.8
	Relevance	24.6	26.4	49.0
	Consistency	2.6	13.2	84.2
VIST-SIS	Visual Quality	14.6	20.6	64.8
	Relevance	19.2	48.6	32.2
	Consistency	3.0	46.2	50.8

Table 3: AR-LDM human evaluation vs ground truth



(a) Auto-regression process.

(b) Model Architecture.

Figure 6: AR-LDM model architecture (Pan, et al., 2022).

One of the most recent works aimed at improving upon AR-LDM is ’Intelligent Grimm’. The researchers behind it, trained a new diffusion model called **StoryGen** and claimed to have significantly improved upon its predecessor in terms of story continuation. They constructed a new dataset called StorySalon which consisted of many more story frames, had a much larger average story length and featured a lot more different categories of characters. They claim that it is a much better-quality dataset which should enable a higher order of models finetuned to a story continuation task (Liu, et al., 2024). The researchers created their own story visualisation model, further improving upon their predecessors' architectures by implementing a novel context encoding solution. They took all previous context frames, added noise and processed them in a separate diffusion model for one step in order to denoise the frame (each frame being conditioned with a corresponding prompt). Afterwards, they extracted this visio-textual context, concatenated it and passed over to the main diffusion model as conditioning. This approach may seem overcomplicated at first glance but an extra denoising step, enhances the temporal positional embedding. Additionally, the main attention layers were split into text and visual ones and summed up at the end. The context encoding for past frames described above feeds into image attention. Thanks to these innovations, Intelligent Grimm outscores previous models both in qualitative consistency metrics as well as quantitative and qualitative image quality metrics.

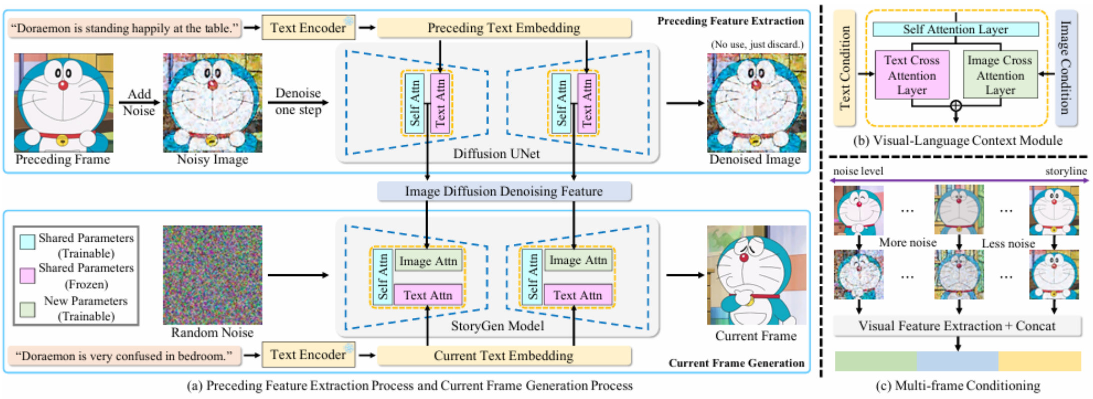
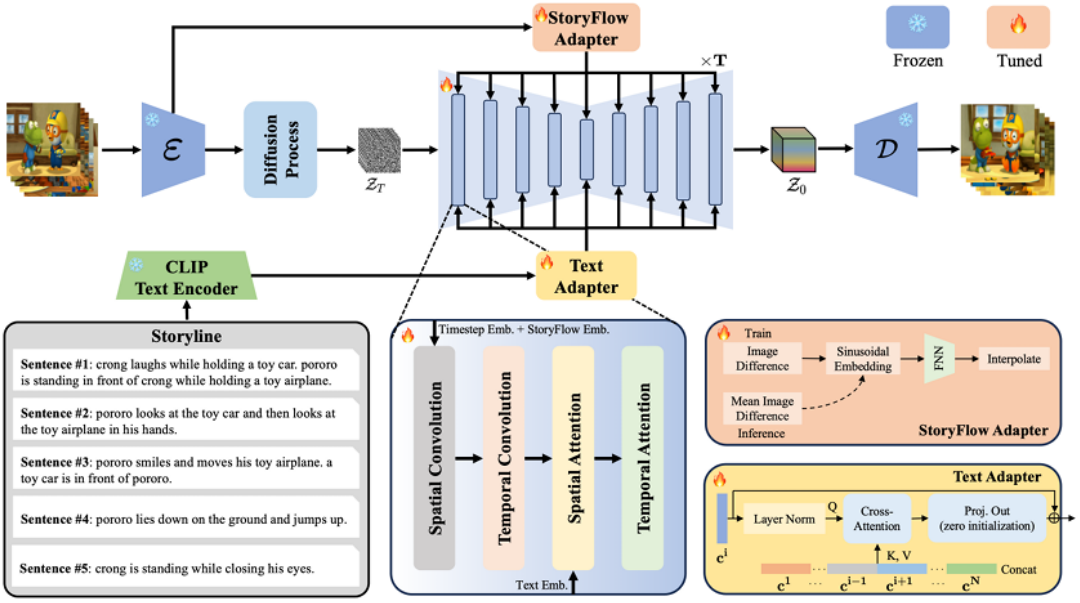

"Doraemon is standing happily at the table."
Text Encoder
Preceding Text Embedding
Preceding Feature Extraction
(No use, just discard.)
Add Noise
Denoise one step
Self Attn
Text Attn
Self Attn
Text Attn
Preceding Frame
Noisy Image
Diffusion UNet
Denoised Image
Image Diffusion Denoising Feature
Shared Parameters (Trainable)
Shared Parameters (Frozen)
New Parameters (Trainable)
Random Noise
Self Attn
Image Attn
Text Attn
Self Attn
Image Attn
Text Attn
StoryGen Model
Current Frame
"Doraemon is very confused in bedroom."
Text Encoder
Current Text Embedding
Current Frame Generation
(a) Preceding Feature Extraction Process and Current Frame Generation Process
Text Condition
Self Attention Layer
Text Cross Attention Layer
Image Cross Attention Layer
Image Condition
(b) Visual-Language Context Module
noise level
storyline
More noise
Less noise
Visual Feature Extraction + Concat
(c) Multi-frame Conditioning

Figure 7: StoryGen model architecture (Liu, et al., 2024).

Most recent advancements have been made by Zheng & Fu in their TemporalStory model. Most notably, they split attention layers into spatial and temporal blocks and introduce spatial and temporal convolution blocks to better capture the dependencies between the frames and therefore improve consistency (Zheng & Fu, 2024). Thanks to these advancements, TemporalStory outscored AR-LDM in user studies (Table 4) and is now considered a state-of-the-art Story Visualisation model using PororoSV and FlintstonesSV datasets.

Dataset	Attribute	Ours	Tie	AR-LDM
PororoSV	Visual Quality	**81.0%**	6.9%	12.1%
	Semantic Relevance	**85.6%**	9.2%	5.2%
	Temporal Consistency	**84.1%**	8.8%	7.1%
FlintstonesSV	Visual Quality	**80.4%**	6.2%	13.4%
	Semantic Relevance	**82.6%**	6.3%	11.1%
	Temporal Consistency	**84.8%**	5.4%	9.8%

Table 4: TemporalStory human evaluation vs AR-LDM



Sentence #1: crong laughs while holding a toy car. pororo is standing in front of crong while holding a toy airplane.

Sentence #2: pororo looks at the toy car and then looks at the toy airplane in his hands.

Sentence #3: pororo smiles and moves his toy airplane. a toy car is in front of pororo.

Sentence #4: pororo lies down on the ground and jumps up.

Sentence #5: crong is standing while closing his eyes.

Figure 8: TemporalStory model architecture (Zheng & Fu, 2024).

“Practical” trope

The most notable difference between the models discussed above and those we will discuss in this section, is the adaptability and flexibility. The following works focus on achieving the results while avoiding costly diffusion model training by building around the diffusion components and creatively manipulating the conditioning network.

Initial solutions to the consistency problem involved a method called **identity injection**. This can be done in many different ways. Perhaps the simplest form of identity injection is **textual inversion** (Gal, et al., 2022) whereby a subject is injected into the network in the form of a text embedding. This was successfully implemented into a consistency mechanism by Jeong, et al. in their work on training-free storybook generation, where the researchers developed a pipeline for generating consistent storybooks from plain text (Jeong, et al., 2023). They used a pre-trained LLM that was guiding a Latent Diffusion Model with a textual inversion module for identity injection to maintain consistency. Their paper focuses heavily on their identity injection algorithm. They compared its results with those from other models like CLIP-guided Diffusion, DALL-E 2 or Blended Latent Diffusion in a set of small-scale questionnaires asking users to assess each model’s correspondence, consistency, and smoothness. The conclusion was that their algorithm outperformed these models although the authors later suggest that not all reference models could perform to their best due to researchers’ limitations which might have led to suboptimal results in some cases.

Another forms of identity injection are so called adapters. One of the landmark papers was written by Ye et al. in 2023. They proposed an **IP-Adapter** model which allows the image generation model to take an additional image as an input. It does so by adding an additional linear layer and Layer Normalization to the standard CLIP image encoder in order to extract image features which are then integrated into the diffusion model with a decoupled cross-attention module alongside the text features from the prompt.

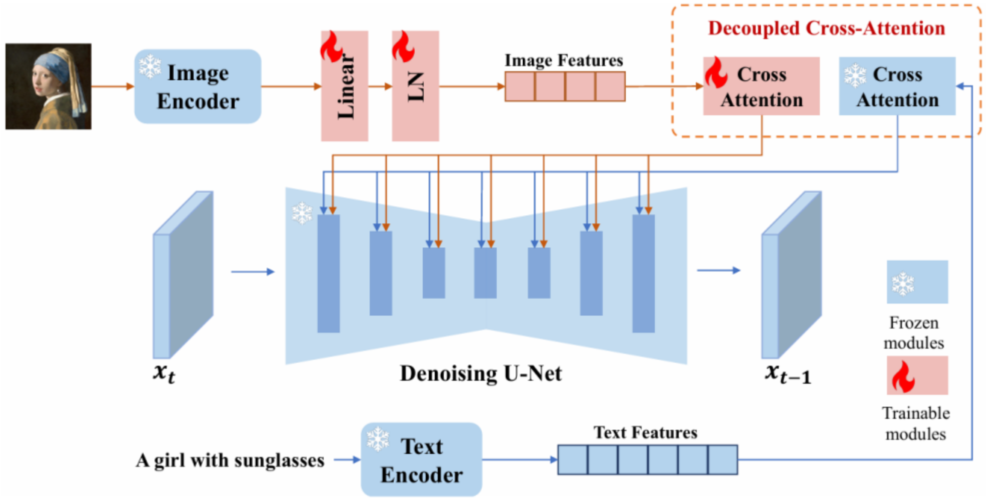


Figure 9: IP-Adapter model architecture (Ye, et al., 2023).

In result, the generated images are a derivative of the reference image. The model is capable of feature injection, style conversion and structural control. However, despite its effectiveness it struggles to maintain the consistency between the generated images. Nevertheless, it is still widely used as a lightweight tool for simple image generation tasks.

Different solution has been presented by Gong et al. in their recent work on the TaleCrafter pipeline (Gong, et al., 2023). They proposed a simple, complete four-stage pipeline for story to video conversion that includes the following modules: S2P (story-to-prompt), T2L (text-to-layout), C-T2I (controllable text-to-image) and I2V (image-to-video). In the original implementation, the authors used GPT-4 (OpenAI, 2023) in their S2P module, LayoutDM (Inoue, et al., 2023) in the T2L module followed by their own diffusion model based on a Stable Diffusion v1.4 (Rombach, et al., 2021).

However, arguably the most interesting work has been published just recently, as a group of researchers from NVIDIA published their work on a training-free consistent image generation called ConsiStory. Tewel et al. came up with a clever solution for preserving the consistency of desired attributes and leaving other attributes unconditioned. They applied a foreground mask to conditioning frames before passing them to the self-attention layer, in effect causing the consistency effect to be isolated only to a foreground of the frame which usually depicts a subject of the frame. They dubbed this technique Subject-Driven Self-Attention (SDSA). The researchers further enhanced the model by introducing a self-attention dropout layer to SDSA. Dropout is a mechanism whereby a random subset of inputs is randomly nullified. It is often being used in neural network training to inject stochasticity to the network which may help alleviate many common problems such as overfitting. In the context of image generation, dropout layer is used to prevent mode collapse (state in which a model keeps producing very similar outputs). The researchers found that dropout rate directly controls the trade-off between consistency and background variation and can be adjusted to strike a balance between the two.

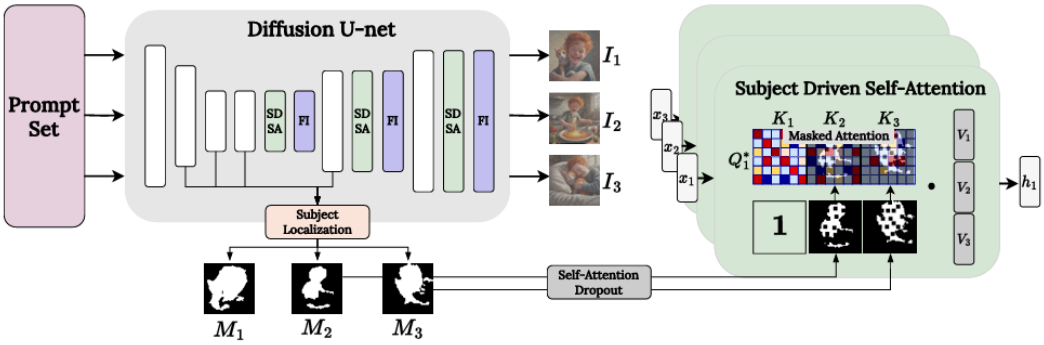


Figure 10: ConsiStory architecture (Tewel, et al., 2024).

These results are further refined by a feature injection algorithm that aims to tackle the finest of details that defines the subject's identity and may get omitted by the model. Finally, the whole algorithm does not require any training and is, theoretically, applicable to any diffusion model (Tewel, et al., 2024).

Sadly, despite solid qualitative results, the researchers fail to not provide enough quantitative results as they only report on the average CLIP-score (metric evaluating image-to-prompt similarity) (Hessel, et al., 2021) and the user study only comparing it to IP-Adapter (Ye, et al., 2023), Textual Inversion (Gal, et al., 2022) and LoRa-related solutions (Hu, et al., 2022). Furthermore, the researchers identified that the feature injection algorithm is prone to failure which according to their observations, results in failures in less than 5% of attempts. This may be due to their methods' weak capabilities to work with different styles at the same time.

Chapter 3: Requirements and Implementation

Quantitative evaluation problem

A keen reader will notice that in the review above, we often referred to user studies. This was because there is no standard and agreed upon metric measuring subject consistency across multiple frames. A lot of the major works summarised above have failed to find or develop a robust consistency evaluation metric. Due to this, up until now all consistency scores are qualitative results obtained via human evaluation. In 'Intelligent Grimm', the researchers even point out the lack of such metric in their justification for their human evaluation studies (Liu, et al., 2024). That is not to say that there aren't any candidates.

Two such metrics that are worth of note are Character F1 score and Character Accuracy. We can see that these metrics are used in StoryDALL-E and TemporalStory papers. They're both based on an InceptionV3 model and can accurately evaluate which characters in the story are portrayed and which are missing. However, they are not good metrics to evaluate the consistency of the characters as they only measure it indirectly. Furthermore, they rely on labels provided by the dataset which renders them impractical for our use case as they require the presence of ground truth data.

The metric that is better suited for this task has been proposed back in 2020 by Song et al. as part of their work surrounding a GAN-based consistency story visualisation model (Song, et al., 2020). The researchers proposed a **Fréchet Story Distance** (**FSD**) metric which is de-facto a modified FVD (Fréchet Video Distance) metric that was adapted to accommodate sequences of images of arbitrary length (Unterthiner, et al., 2018). Specifically, they swapped out the I3D (Two-Stream Inflated 3D ConvNets) network (Carreira & Zisserman, 2017) that was used by the FVD metric for a more suitable R(2+1)D network (Tran, et al., 2018). Both are video action classification networks – the former is a fully spatiotemporal network while the latter is an improved variant which separated the spatial and temporal components into two serialised blocks of 2D and 1D convolutions. Despite being widely cited, this aspect of Song et al.'s work has gone largely unnoticed over the years. In the 'Results' section, we will revisit this metric and attempt to evaluate the consistency of some of the major models described in the Literature Review.

However, FSD does not solve the main research question of this project. Being a derivative of the FID metric, it relies on ground truth data and therefore it cannot be used to evaluate the consistency of a general image generation model where there is no ground truth. We are still yet to bridge the gap between the two trails of thought described in the literature review. To address this, we introduce a new **Fréchet Subject Similarity Distance** (**F2SD**) evaluation metric. It substitutes the ground truth reference frames for a single reference frame depicting the subject of our story. The F2SD is a natural progression of FSD which only concerns generated frames making it applicable to all families of story visualisation models and focuses purely on the consistency between the source frame and the target frames. Furthermore, **F2SD** metric uses frames with 256x256 resolution as opposed FSD's 64x64.

To summarize:

Fréchet Story Distance – measures a Fréchet distance between a sequence of generated frames and a sequence of reference frames of the same length. Introduced by Song et al. in 2020.

Fréchet Subject Similarity Distance – a new metric that measures a Fréchet distance between each of the series of frames and a reference frame.

Fréchet Distance

Fréchet distance between two sequences of frames P_R and P_G is a 2-Wasserstein distance which is a difference between the sum of the mean squared error and the traces of both covariance matrices, and a trace of the geometric mean of those covariance matrices.

It can be expressed as follows:

$$d_{FD}(P_R, P_G) = |\mu_R - \mu_G|^2 + Tr\left(\Sigma_R + \Sigma_G - 2\sqrt{(\Sigma_R\Sigma_G)}\right),$$

where μ_R and μ_G are means and Σ_R and Σ_G are covariance matrices of the data distribution acquired from the last average pooling layer's output of the evaluation model (Song, et al., 2020). FSD and FID vary only in which model they are using to acquire the distributions. F2SD uses just a single reference frame, therefore it can be expressed as:

$$d_{F2SD}(p_R, P_G) = d_{FD}(\{p_R^n\}, P_G),$$

where p_R turns into a multiset of length n, equal to the length of P_G.

Implementation

The Python implementation is reusing the FID score evaluation built by Heusel et al. and ported to PyTorch by Maximilian Seitzer as well as the Song et al.'s implementation of FSD (Heusel, et al., 2017), (Seitzer, 2020), (Song, et al., 2020).

Full source code can be accessed here: https://github.com/cwirzenm/F2SD.

Chapter 4: Methodology

The methodology applied in this project is based on four main principles: design, research, implementation, and testing. These principles are assisted by the system of daily and weekly goals, that are to be achieved. It is a modification of the Waterfall methodology that was simplified and adapted to a single person led project. It is a comfortable and solid choice due to its past successes in other ML projects.

This methodology predicts sudden changes in design, setbacks and necessary pivots. The first design stage should outline the main objective of the project and describe in little detail how to get to that objective. After that, the user should go in more and more detail with each iteration. During the research stage, a user should gain the practical and theoretical knowledge about how to complete the task that was described in the design stage. This is followed by the implementation stage, where this knowledge should be applied. The testing stage blends in with the implementation stage as it is a part of the practical work, however it goes more in depth into the different variables and configurations of the system which is particularly crucial in AI applications.

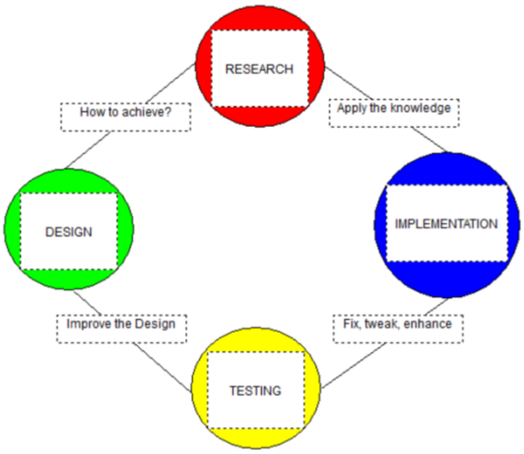


Figure 11: Applied methodology diagram

Applying the methodology outlined above, the first cycle focused on experimentation with various pipelines involving background subtraction. It quickly became apparent that unless using a particularly robust model, it is not going to work as we found that this solution works very well only with realistic or close to realistic styles of frames. Due to this fact, the idea of background subtraction was dropped. The second cycle, pivoted towards object detection. It was found that the YOLOV8x model (Dong, et al., 2024) works particularly well with all frames, regardless of style. However, the F2SD evaluation using YOLOV8x proved to be very inconsistent and unreliable. That was most likely because the model embeddings contained too much information that was irrelevant from the consistency perspective such as object identification. It should be noted that this approach appears to be the most promising way forward, that sadly falls beyond the capabilities and resource requirements of this project. In the end the solution settled at **F2SD** paired with the much simpler R(2+1)D model.

The testing data that was used throughout this project was acquired from the StoryDALL-E (Maharana, et al., 2022), AR-LDM (Pan, et al., 2022), TemporalStory (Zheng & Fu, 2024) and ConsiStory (Tewel, et al., 2024) works. It was manually processed to become usable for the purposes of this project (split into frames and sorted into ‘stories’). The testing was deemed successful once the data reliably showed clear distinction between the story frames and the noisy frames across multiple different configurations.

Chapter 5: Results

Model	number of samples					
	Pororo	Poby	Harry	Fred	Wilma	Barney
StoryDALL-E (Maharana, et al., 2022)	5	5	0	6	7	0
AR-LDM (Pan, et al., 2022)	10	10	8	10	10	8
TemporalStory (Zheng & Fu, 2024)	9	8	5	10	9	9

Table 5: Number of samples from each dataset for F2SD evaluation

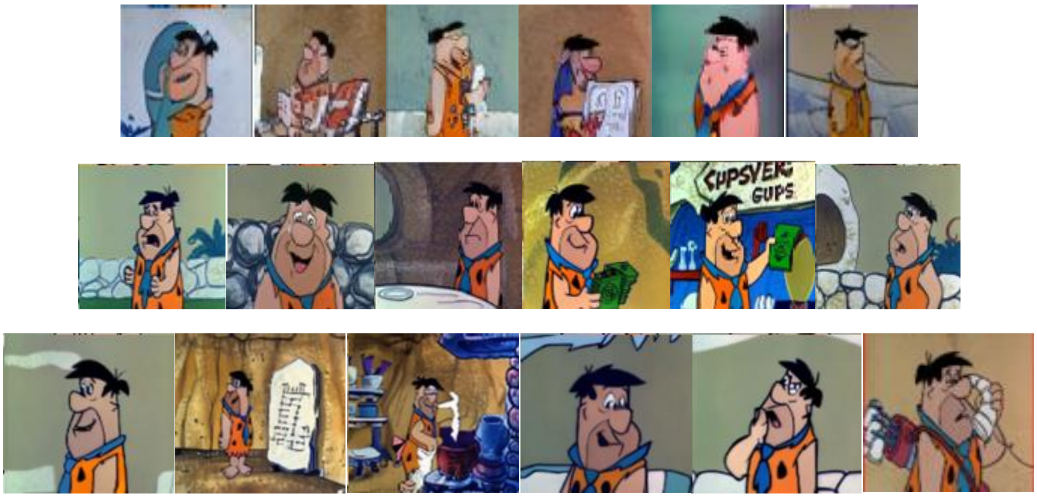


Figure 12: 'Fred' test samples from different datasets. From top to bottom: StoryDALL-E, AR-LDM, TemporalStory

Model	PororoSV			FlintstonesSV		
	FID↓	FSD↓	Avg. F2SD↓	FID↓	FSD↓	Avg. F2SD↓
StoryDALL-E (Maharana, et al., 2022)	25.90	471.34	81.03	26.49	489.90	**54.66**
AR-LDM (Pan, et al., 2022)	17.40	**200.60**	77.26	19.28	223.30	78.68
TemporalStory (Zheng & Fu, 2024)	**14.20**	284.13	**66.44**	**16.33**	**321.85**	69.56

Table 6: FID, FSD and F2SD scores by dataset

Model	FlintstonesSV F2SD↓		
	Fred	Wilma	Barney
StoryDALL-E (Maharana, et al., 2022)	60.38	**48.93**	-
AR-LDM (Pan, et al., 2022)	75.84	94.24	65.97
TemporalStory (Zheng & Fu, 2024)	**58.66**	85.73	**64.30**

Model	PororoSV F2SD↓		
	Pororo	Poby	Harry
StoryDALL-E (Maharana, et al., 2022)	62.17	99.89	-
AR-LDM (Pan, et al., 2022)	90.77	86.98	**54.04**
TemporalStory (Zheng & Fu, 2024)	**56.70**	**70.23**	72.38

Table 7 & 8: Detailed F2SD scores by dataset

Model	Avg. F2SD↓
StoryDALL-E (Maharana, et al., 2022)	**67.84**
AR-LDM (Pan, et al., 2022)	77.97
TemporalStory (Zheng & Fu, 2024)	68.00
ConsiStory (Tewel, et al., 2024)	86.31
Noise	128.28

Table 9: Average F2SD scores across all domains

Table 6 presents the average **F2SD** results of the most significant works in the field and compares them side-by-side with their corresponding FID and FSD scores. Tables 7 & 8 provide more insight into scores achieved by frames representing each character in the story. We can see that according to F2SD, StoryDALL-E performs surprisingly well on FlintstonesSV. It can be explained by noting that it achieved by far the worst FID and FSD scores, meaning that it wasn't able to convey as many details compared to the ground truth image. Due to the fact that, F2SD does not take ground truth images into account, we can see how that result would have been expected. In other words, the model was consistent in generating poorly approximated Flintstones story frames. On PororoSV, the StoryDALL-E wasn't as consistent which could be attributed to fewer available samples. Yet it wasn't far off from the more sophisticated AR-LDM and TemporalStory models which consistently scored around 77 and 67 respectively. Sample size is an important factor which ought not to be overlooked. Some configurations had less frames than others. This encourages higher variation and suggests that further testing is required in order to achieve more reliable results.

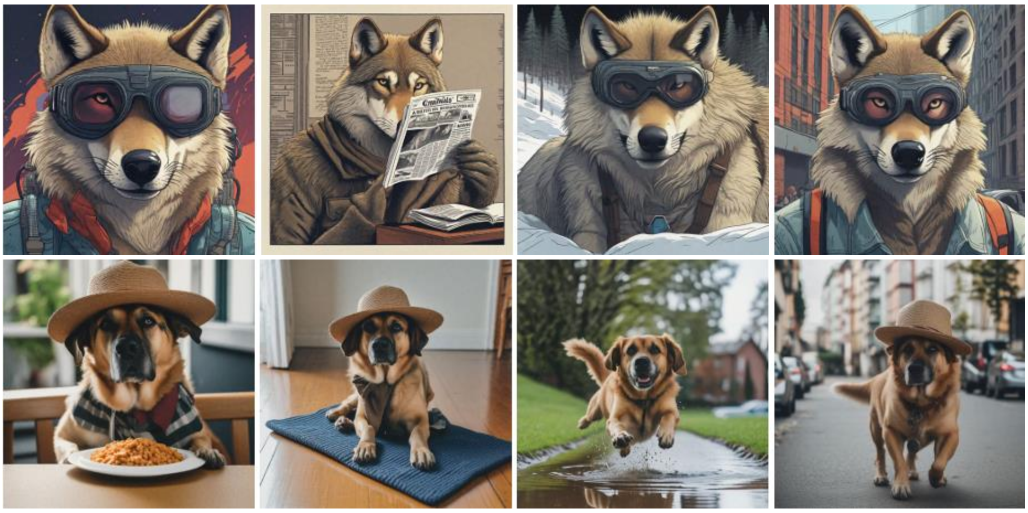


Figure 13: Test samples from ConsiStory

Next, when we look at Table 9, we can see how ConsiStory compares with the rest of the baseline models. From the results we can deduce that it might be the case opposite to that of StoryDALL-E. When looking at the sample frames (Figure 13), it's evident that the quality of the images generated by ConsiStory far exceeds those by other models - therefore, the model has to keep track of more features, which explains the worse consistency score. Even with those caveats, the F2SD metric succeeds in putting the models from two different approaches on the same level. It’s worth noting that the ConsiStory dataset consisted of 88 images making up 20 ‘stories’ with the final score being an average of all ‘story’ scores.

Finally, notice the ‘Noise’ score in Table 9. This dataset consisted of 25 frames that were randomly picked from all other datasets. The goal of this dataset was to provide a reference to what a bad score would look like. It is important to highlight that the final scores highly depend on the resolution of the inputted frames. It was found that the higher the resolution, the larger the scores get.

Chapter 6: Conclusions and Future Work

The objective of this project was to level the playing field in consistent image generation between the models that utilise objective ground truth data and those that don’t. We demonstrate that the F2SD evaluation metric achieves this goal by relying purely on generated data, making it far more flexible than its predecessor – FSD. The acquired results suggest that F2SD only tells part of the story and when evaluating the results of a model, one should consider the quality of the images compared. Naturally, if one was to compare models that use the same diffusion model as backbone, standalone F2SD becomes more viable.

Perhaps one of the most promising avenues for future development would be to use a more robust network to evaluate the Fréchet Distance. R(2+1)D network has been introduced long before the image generation was robust and impressive. Nowadays, the models are far more sophisticated and advanced, and it would be an interesting study to see if it’s possible to get more detailed Fréchet Distance scores. This project attempted to swap the network for far more robust YOLOv8x (Dong, et al., 2024), but this endeavour ended with a conclusion that the embeddings coming from such complex network would have to be further processed.

Furthermore, it should be investigated whether R(2+1)D suffers from the same deficiency as the Inception model does (used originally in FID metric). Recently, Jayasumana et al. demonstrated that the Inception model suffers from a deficiency whereby it inherently assumes that all data is normally distributed, which may lead to some false positives. They proposed a CLIP-based metric called MMD distance (Maximum Mean Discrepancy) which alleviates the issue (Jayasumana, et al., 2024). Given the above findings, this appears to be an area worth exploring further.

Another interesting prospect would be to merge the consistency score with an image quality metric. In this study, it was found that the results must be interpreted carefully due to its detachment from the quality, where worse quality images achieved higher scores as they had less details to convey onto other images. Perhaps a metric which would take that into consideration, would be more intuitive and user-friendly.

References

Abdul-Rahman, A. et al., 2013. Rule-based Visual Mappings –with a Case Study on Poetry Visualization. *Computer Graphics Forum,* 32(3), pp. 371-490.

Ansag, R. A. & Gonzalez, A. J., 2021. State-of-the-Art in Automated Story Generation Systems Research. *JOURNAL OF EXPERIMENTAL & THEORETICAL ARTIFICIAL INTELLIGENCE,* 35(6), pp. 877-931.

Bahdanau, D., Cho, K. & Bengio, Y., 2014. *Neural Machine Translation by Jointly Learning to Align and Translate.* s.l., s.n.

Bajrami, D. & Richters, C., 2023. *AI Storyboard Generator – Script Forge,* Saxion: Saxion University of Applied Sciences.

boords, 2023. *AI Storyboard Generator - Script to Storyboard with AI | Boords.* [Online]
Available at: https://boords.com/ai-storyboard-generator
[Accessed 2023].

Bottoni, B. et al., 2020. *Character Depth and Sentence Diversification in Automated Narrative Generation.* North Miami Beach, FL, AAAI Press.

Carreira, J. & Zisserman, A., 2017. *Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset.* Honolulu, HI, IEEE.

DeepLearning.AI, 2023. *Natural Language Processing (NLP) [A Complete Guide].* [Online]
Available at: https://www.deeplearning.ai/resources/natural-language-processing/
[Accessed December 2024].

Dhariwal, P. & Nichol, A., 2021. Diffusion Models Beat GANs on Image Synthesis.

Dong, C., Tang, Y., Zhu, H. & Zhang, L., 2024. HCA-YOLO: a non-salient object detection method based on hierarchical attention mechanism. *Cluster Computing,* 25 April, Volume 27, p. 9663–9678.

Gal, R. et al., 2022. *An Image is Worth One Word: Personalizing Text-to-Image Generation using Textual Inversion,* s.l.: s.n.

Gong, Y. et al., 2023. *TaleCrafter: Interactive Story Visualization with Multiple Characters.* Sydney, Association for Computing Machinery.

Hariffadzillah, T. S. N. T. et al., 2023. *Exploring Ai Technology To Create Children's.* IPOH, Malaysia, IEEE.

Hessel, J. et al., 2021. CLIPScore: A Reference-free Evaluation Metric for Image Captioning. *CoRR.*

Heusel, M. et al., 2017. *GANs trained by a two time-scale update rule converge to a local nash equilibrium.* Long Beach, CA, USA, Curran Associates, pp. 6629-6640.

Hollister, J. R. & Gonzalez, A. J., 2018. The campfire storytelling system – automatic creation and modification of a narrative. *Journal of Experimental & Theoretical Artificial Intelligence,* 31(1), pp. 15-40.

Hu, E. J. et al., 2022. *LoRA: Low-Rank Adaptation of Large Language Models.* s.l., ICLR.

huggingface, 2023. *Textual Inversion.* [Online]
Available at: https://huggingface.co/docs/diffusers/training/text_inversion
[Accessed December 2023].

huggingface, n.d. *Diffusers.* [Online]
Available at: https://huggingface.co/docs/diffusers/index
[Accessed 2023].

Inoue, N. et al., 2023. *LayoutDM: Discrete Diffusion Model for Controllable Layout Generation.* Vancouver, BC, Canada, IEEE, pp. 10167-10176.

Jayasumana, S. et al., 2024. *Rethinking FID: Towards a Better Evaluation Metric for Image Generation.* s.l., s.n.

Jeong, H., Kwon, G. & Ye, J. C., 2023. Zero-shot Generation of Coherent Storybook from Plain Text Story using Diffusion Models. *Computer Vision and Pattern Recognition,* 8 February.

Kahn, K. M., 1979. *Creation of computer animation from story descriptions.,* s.l.: s.n.

Karadoğan, A. & Uyumaz, F., 2023. Visualisation of Some Poems of Yahya Kemal with Artificial Intelligence Technology. *International Journal of Textbooks and Education Materials,* 6(2), pp. 211-235.

krock.io, 2023. *AI Storyboard Generator - Storyboarder AI | KROCK.IO.* [Online]
Available at: https://krock.io/storyboard-ai/
[Accessed December 2023].

Kuprel, B., 2022. *kuprel/min-dalle.* [Online]
Available at: https://github.com/kuprel/min-dalle

Lebowitz, M., 1985. Story-telling as planning and learning.. *Poetics,* 14(6), pp. 483-502.

Lian, L., Li, B., Yala, A. & Darrell, T., 2023. LLM-grounded Diffusion: Enhancing Prompt Understanding of Text-to-Image Diffusion Models with Large Language Models. *arXiv preprint arXiv:2305.13655.*

Li, J., Li, D., Xiong, C. & Hoi, S., 2022. *BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation.* s.l., PMLR, pp. 12888-12900.

Ling, Y. et al., 2023. Diffusion Models: A Comprehensive Survey of Methods and Applications. *ACM Computing Surveys,* 9 November, 56(4), pp. 1-39.

Liu, C. et al., 2024. *Intelligent Grimm -- Open-ended Visual Storytelling via Latent Diffusion Models.* Seattle, IEEE, pp. 6190-6200.

Maharana, A., Hannan, D. & Bansal, M., 2022. *StoryDALL-E: Adapting Pretrained Text-to-Image Transformers for Story Continuation.* s.l., Springer, p. 70–87.

Meehan, J. R., 1977. *TALE-SPIN: An Interactive Program that Writes Stories..* Cambridge, MA, Morgan Kaufmann Publishers Inc., pp. 91-98.

Meta, 2023. *Llama 2: Open Foundation and Fine-Tuned Chat Models.* [Online]
Available at: https://arxiv.org/abs/2307.09288

Microsoft, 2023. *Phi-2: The surprising power of small language models.* [Online]
Available at: https://www.microsoft.com/en-us/research/blog/phi-2-the-surprising-power-of-small-language-models/
[Accessed December 2023].

MidJourney, 2022. *MidJourney.* [Online]
Available at: https://www.midjourney.com/home?callbackUrl=%2Fexplore
[Accessed 2023].

MidJourney, 2023. *Midjourney v6.* [Online]
Available at: https://mid-journey.ai/midjourney-v6-release/
[Accessed December 2023].

Mistral, 2023. *Mixtral of experts.* [Online]
Available at: https://mistral.ai/news/mixtral-of-experts/
[Accessed December 2023].

ModelScope, 2023. *damo-vilab/text-to-video-ms-1.7b.* [Online]
Available at: https://huggingface.co/damo-vilab/text-to-video-ms-1.7b
[Accessed December 2023].

National Geographic Society, 2023. *Storytelling.* [Online]
Available at: https://education.nationalgeographic.org/resource/storytelling-x/
[Accessed 26 November 2024].

Natural Synthetics Inc., 2023. *hotshotco/Hotshot-XL.* [Online]
Available at: https://huggingface.co/hotshotco/Hotshot-XL
[Accessed December 2023].

OpenAI, 2023. *DALL·E 3 System Card.* [Online]
Available at: https://openai.com/research/dall-e-3-system-card

OpenAI, 2023. *GPT-4 Technical Report,* s.l.: s.n.

Pan, X. et al., 2022. *Synthesizing Coherent Story with Auto-Regressive Latent Diffusion Models.* s.l., The Computer Vision Foundation.

Peebles, W. & Xie, S., 2023. *Scalable Diffusion Models with Transformers.* Paris, IEEE, pp. 4195-4205.

Radford, A. et al., 2021. *Learning Transferable Visual Models From Natural Language Supervision.* s.l., PMLR , pp. 8748-8763.

Rombach, R. et al., 2021. *High-Resolution Image Synthesis With Latent Diffusion Models.* s.l., Computer Vision Foundation, pp. 10684-10695.

Seitzer, M., 2020. *pytorch-fid: FID Score for PyTorch,* s.l.: s.n.

Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N. & Ganguli, S., 2015. *Deep Unsupervised Learning using Nonequilibrium Thermodynamics.* Lille, France, s.n., pp. 2256-2265.

Song, Y.-Z.et al., 2020. *Character-Preserving Coherent Story Visualization.* Glasgow, Springer, pp. 18-33.

Stability AI, 2023. *stabilityai/stable-video-diffusion-img2vid-xt.* [Online]
Available at: https://huggingface.co/stabilityai/stable-video-diffusion-img2vid-xt
[Accessed December 2023].

StoryboardHero, 2023. *AI Storyboard Generator - Free Trial - Storyboard Hero.* [Online]
Available at: https://storyboardhero.ai/
[Accessed 2023].

storybookai, 2023. *StoryBook AI | The AI-Powered Story Generator.* [Online]
Available at: https://www.storybookai.app/

Tewel, Y. et al., 2024. *Training-Free Consistent Text-to-Image Generation,* s.l.: s.n.

Tolstoy, L., 1869. *War and Peace.* s.l.:s.n.

Tran, D. et al., 2018. *A Closer Look at Spatiotemporal Convolutions for Action Recognition.* Salt Lake City, UT, IEEE.

Unterthiner, T. et al., 2018. *Towards accurate generative models of video: A new metric & challenges.,* s.l.: s.n.

Vaswani, A. et al., 2017. *Attention Is All You Need.* Long Beach, CA, Google.

Wade, J. et al., 2017. *A Stochastic Approach to Character Growth in Automated Narrative Generation.* Marco Island, FL, AAAI Press.

Wu, J. Z. et al., 2023. *Tune-A-Video: One-Shot Tuning of Image Diffusion Models for Text-to-Video Generation.* s.l., IEEE, pp. 7623-7633.

Ye, H. et al., 2023. *IP-Adapter: Text Compatible Image Prompt Adapter for Text-to-Image Diffusion Models.* s.l., s.n.

Zheng, S. & Fu, Y., 2024. *TemporalStory: Enhancing Consistency in Story Visualization using Spatial-Temporal Attention.* s.l., s.n.