Evaluating Visual Narrative Coherence in Story Visualization via Diversified Storylines

Minha Jhang[1] **Kyeongman Park** [1] **Hyukhun Koh**[1] **Kyomin Jung**[1,2]
[1]IPAI, Seoul National University
[2]Dept. of ECE, Seoul National University
{jminha2014, zzangmane, hyukhunkoh-ai, kjung}@snu.ac.kr

Abstract

Story visualization requires generating a coherent sequence of images that collectively form a narrative, yet existing evaluation metrics and datasets often overlook visual continuity and narrative diversity. In this paper, we introduce the *Visual Context-Aware Metric for Story Visualization (VCMS)*, which uses large vision-language models to jointly assess caption fidelity and inter-image consistency, achieving Spearman's correlation comparable to human agreement on two benchmarks. Also, to address the shortcomings of narrowly defined datasets with low diversity, we propose a *diffusion-augmented evaluation pipeline* that blends diverse and controlled narrative elements at adjustable ratios, enabling the creation of evaluation sets tailored to specific objectives. By combining VCMS with this pipeline, we provide a scalable, human-aligned framework for evaluating story visualization models.

1 Introduction

Story visualization demands generating a coherent image sequence that forms a unified visual narrative; however, current evaluation metrics and datasets frequently neglect *visual continuity* and *narrative diversity*. To remedy this, more comprehensive evaluation metrics and a controllable, diverse evaluation set are required.

As a challenging subfield of text-to-image generation, story visualization demands that each image not only accurately reflects its caption but also maintains consistent characters, backgrounds, and stylistic elements throughout the sequence(Liu et al., 2024; Yang and Jin, 2024). However, evaluation metrics such as CLIP-based scores (Hessel et al., 2021) fail to capture the nuanced narrative coherence and visual continuity required for high-quality story visualization. Moreover, conventional closed-ended evaluation datasets—typically featuring fixed storyline styles (Liu et al., 2024)—do not reflect the open-ended adaptability demanded by real-world storytelling, where vocabulary, style, and context vary, and they cannot effectively distinguish models overfitted to closed narratives, as demonstrated in our experiments.

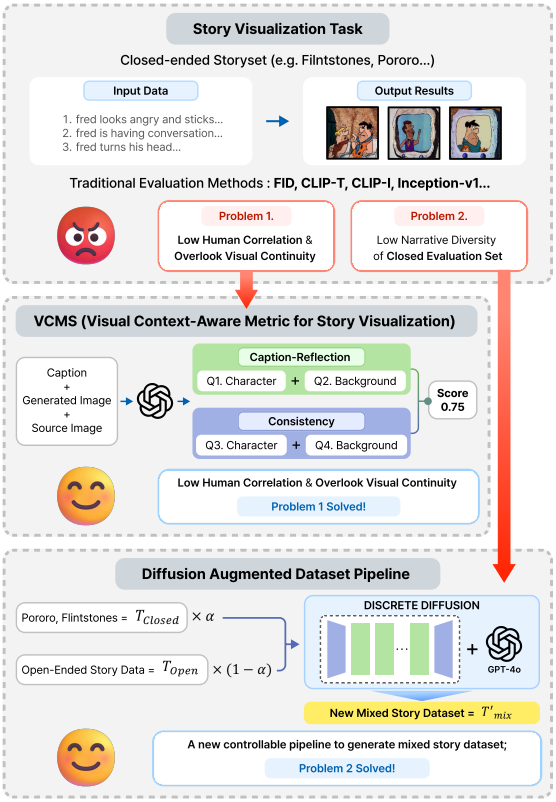

Figure 1: Overview of our approach towards fine-grained and diversified story visualization evaluation.

To address these limitations of *fragmented expression* and *restricted evaluation*, we introduce two novel methods:

- **Visual Context-Aware Metric for Story Visualization (VCMS):** A large vision language model (LVLM)-based evaluation framework that jointly assesses caption fidelity and inter-image consistency.
- **Diffusion-Augmented Dataset Generation**

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 34192–34207

Pipeline: A pipeline to create a diverse mixed dataset T'_{mixed} that blends closed-ended and open-ended narratives for enhanced diversity and continuity.

First, we present the Visual Context-Aware Metric for Story Visualization (VCMS), an evaluation framework that leverages LVLM such as GPT-4o (OpenAI, 2023). VCMS jointly assesses caption fidelity and inter-image consistency by posing carefully designed prompts. This approach quantifies how well each generated image adheres to its caption and successive images preserve visual context. To our knowledge, it is the first framework to employ LVLM-based evaluation specifically for story visualization, capturing neglected nuances and aligning with human evaluation.

Second, we propose a diffusion-augmented dataset generation pipeline to overcome the limitations of conventional datasets. Our pipeline creates a mixed dataset T_{mixed} by blending fixed narrative elements with open-ended storytelling, controlled by a mix ratio α. This design preserves key narrative structures while introducing the variability needed for robust evaluation. We then use this dataset to train a discrete diffusion language model (Koh et al., 2024a) that yields diverse outputs. Compared to autoregressive LLMs, our approach offers finer control over narrative diversity and style, reducing the need for extensive prompt engineering and mitigating pre-trained bias issues (Chen and Yang, 2023; Feng et al., 2023; Koh et al., 2024b).

Our experiments show that VCMS achieves 95% of the Spearman's rank correlation observed between human annotators, surpassing both traditional and LVLM-based metrics. Furthermore, the diffusion-augmented evaluation set reveals that while some models excel under conventional settings, pre-trained models adapt more effectively to diverse, open-ended narratives. Collectively, our contributions provide a *scalable, human-aligned framework* that *highlights current models' strengths and limitations*, paving the way for future advancements in coherent visual storytelling.

2 Preliminaries: Story Visualization

2.1 Defining Story Visualization

Story visualization differs from standard text-to-image tasks by requiring the generation of coherent image sequences that form a unified narrative. Each image must not only reflect its corresponding text description but also contribute to an overall connected storyline(Yang and Jin, 2024). To formalize the discussion, we define the complete sets of vocabulary, situations, and characters available from various sources as $\mathcal{V}$, $\mathcal{S}$, and $\mathcal{C}$, respectively. In practice, different approaches have been used to construct story visualization datasets. For example, a **closed-ended** approach (Gupta et al., 2018; Maharana et al., 2022; Li et al., 2019) restricts the dataset to limited subsets—denoted as $\mathcal{V}_{\text{closed}} \subset \mathcal{V}$, $\mathcal{S}_{\text{closed}} \subset \mathcal{S}$, and $\mathcal{C}_{\text{closed}} \subset \mathcal{C}$—which simplifies training but limits narrative diversity and style adaptation. In contrast, open-ended approaches, such as that introduced in StorySalon (Liu et al., 2024), leverage the full sets $\mathcal{V}$, $\mathcal{S}$, and $\mathcal{C}$ to support more diverse and flexible narratives. Nevertheless, because closed-ended benchmarks often use repetitive storylines (e.g., from animation series), they may inadvertently boost coherence and visualization performance by overfitting to specific settings. To address these challenges, Section 4 proposes an evaluation pipeline that utilizes open-ended storylines to construct a storyline-independent set for a more objective assessment of visual continuity.

2.2 Evaluation Metrics for Story Visualization

Evaluating story visualization requires assessing both text-to-image alignment and image-to-image coherence. Common metrics include CLIP-T (Hessel et al., 2021) for measuring alignment, alongside FID (Heusel et al., 2018), CLIP-I, and Inception-v3 F1 scores (Maharana et al., 2021; Szegedy et al., 2015) for coherence. However, these automatic metrics are often supplemented with human evaluations to capture nuances in narrative consistency.

Recent research (Ku et al., 2024) has shown that large vision-language models (LVLMs) can evaluate images with human-level accuracy, effectively assessing elements such as character depiction, background fidelity, and overall visual coherence. Motivated by these findings, Section 3 introduces an LVLM-based evaluation framework designed to replace human judgment, enabling scalable and objective assessments of both visualization quality and visual continuity in story visualization tasks.

3 VCMS: Visual Context-Aware Metric for Story Visualization

We propose VCMS, an evaluation method that leverages LVLMs to jointly assess two key aspects of story visualization:

- **Caption Reflection:** How well the generated image reproduces the character's behavior, emotions, and background details described in the caption.
- **Image Consistency:** The visual continuity between successive images, focusing on consistent character appearance and background style.

Theoretical Motivation Previous works in story visualization have typically focused on either prompt reproduction (e.g., using CLIP scores in SEED-Story (Yang and Jin, 2024) and StoryDALL-E (Maharana et al., 2022)) or overall image quality (e.g., FID). Other studies (e.g., Make-A-Story (Rahman et al., 2023), StoryGAN (Li et al., 2019)) highlight the importance of maintaining consistency in characters and backgrounds. In contrast, VCMS integrates both aspects in a unified framework. Unlike related LVLM-based systems such as VIEScore (Ku et al., 2024), AutoEval-Video (Chen et al., 2024), and VQAScore (Lin et al., 2024)—which target single images or videos—we employ LVLM for robust performance in evaluating complex narrative structures.

Evaluation Setup Each generated story image is evaluated using four targeted questions. For Caption Reflection, we ask Q1: "How accurately does the character's emotions and behavior in the generated image reflect those described in the caption?" and Q2: "How appropriate is the background setting in the generated image compared to what is described in the caption?" For Image Consistency, we ask Q3: "How well does the character's appearance in the generated image match their appearance in the previous scene?" and Q4: "How well does the background in the generated image maintain a consistent artistic style compared to the previous scene?" For each question, we prepend the prompt with "On a scale from 0 to 1," and append strict instructions (e.g., "Do not add any explanations...") to ensure that the responds solely with a numerical score. Additional guidance is provided for certain questions (e.g., emphasizing facial expressions and body language for Q1, or details like clothing and color schemes for Q3 and Q4). Full input prompt is provided in Appendix A.

Scoring and Variants The final VCMS score is computed as the average of the four question scores. We also explore two sampling variants: one by sampling three scores per question (with temperature=1) and averaging them, and another by modifying the final instruction to "Just answer Yes or No," then using the word probabilities—assigning a negative weight to "No"—as the score (Fu et al., 2023). This integrated approach enables VCMS to comprehensively capture both text-caption alignment and image coherence, offering a theoretically grounded evaluation framework that addresses the limitations of prior metrics.

4 Evaluation Framework for Story Visualization with Diffusion-Augmented Datasets

Figure 2 illustrates our proposed evaluation pipeline, which addresses the limitations of existing closed-ended datasets.

Current evaluation sets often rely on predefined storylines, which can mask the shortcomings of models that overfit these limited narrative styles. As a result, such eval sets may fail to reveal issues in inter-image consistency, leading to deceptively high **MAUVE** scores (Pillutla et al., 2023) (a distributional distance metric between two datasets) between the training and evaluation sets, as shown in Table 1.

	PororoSV	FlintstonesSV
Original Eval set	0.4096	0.7875
Augmented_mix_25	0.1163	0.0961
Augmented_mix_50	0.1645	0.0807
Augmented_mix_75	0.0635	0.0587

Table 1: MAUVE scores comparing the evaluation sets (original and augmented with different mix ratios) to their respective training sets for PororoSV and FlintstonesSV. Lower scores indicate a greater divergence between the evaluation and training distributions.

Our pipeline overcomes these limitations by integrating open-ended and closed-ended elements, preserving key characters and settings from the closed-ended domain while leveraging diverse vocabulary and scenarios from open-ended data. Moreover, we can adjust the ratio of open-ended to closed-ended data to control the degree of *"openness"*, thereby enabling a more rigorous evaluation of generation performance across various vocabularies and situations.

We employ a Discrete Diffusion Language Model (DDLM) rather than autoregressive LLMs such as GPT-4 because DDLMs offer greater flexibility in controlling output diversity, noise levels, and the ratio of open-ended to closed-ended data

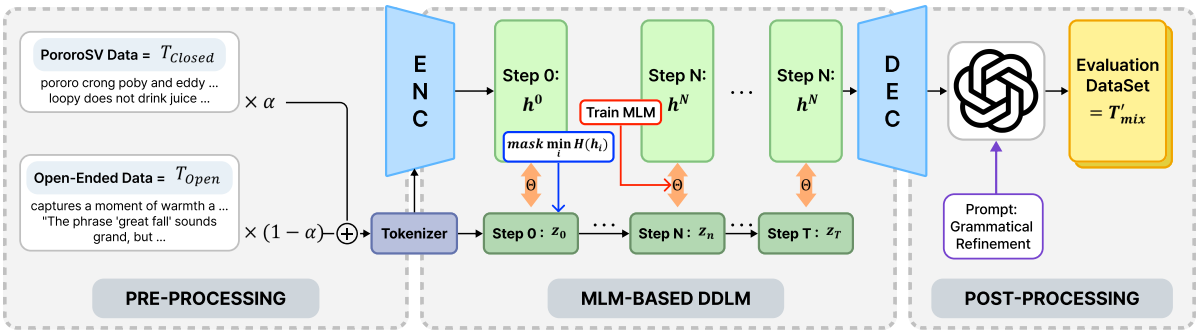

Figure 2: Overview of the data generation pipeline used to create an objective evaluation set. Open-ended and closed-ended texts are blended at a chosen ratio, processed by a discrete diffusion language model (DDLM), and refined with grammatical post-processing, resulting in diverse, coherent storylines.

(Chen and Yang, 2023). In contrast, prompt-tuned or fine-tuned LLMs often require extensive engineering and may be restricted by pretrained biases (Feng et al., 2023; Koh et al., 2024b). By leveraging DDLMs, our pipeline ensures a scalable and adjustable evaluation framework that more accurately reflects real-world storytelling scenarios. A detailed explanation is provided in Appendix I.

4.1 Diversified Narrative Evaluation Structure

To overcome the limitations of closed-ended story visualization datasets, we propose a benchmark pipeline that generates diverse scenarios by blending closed-ended and open-ended data via diffusion language models. Specifically, we augment evaluation diversity by incorporating additional text sequences T_{open}, derived from the complementary sets $\overline{\mathcal{V}_{\text{closed}}}$, $\overline{\mathcal{S}_{\text{closed}}}$, and $\overline{\mathcal{C}_{\text{closed}}}$ (the complements of the closed-ended vocabulary, situations, and characters, respectively). This setup preserves the core characters and settings of the original closed-ended dataset while exposing the model to a broader range of vocabulary, scenarios, and characters, enhancing its generalization in open-ended contexts.

For flexibility, we introduce a **mix ratio** α (where $0 \leq \alpha \leq 1$) to control the proportion of open-ended data in the mixed dataset. Specifically, we sample text sequences T from both the closed-ended set:

$$T_{\text{closed}} \subset \mathcal{V}_{\text{closed}} \times \mathcal{S}_{\text{closed}} \times \mathcal{C}_{\text{closed}},$$

And the open-ended complement set:

$$T_{\text{open}} \subset \overline{\mathcal{V}_{\text{closed}}} \times \overline{\mathcal{S}_{\text{closed}}} \times \overline{\mathcal{C}_{\text{closed}}},$$

to create a mixed dataset T_{mix} defined as:

$$T_{\text{mix}} = (1-\alpha) \cdot T_{\text{closed}} \cup \alpha \cdot T_{\text{open}}.$$

Thus, α precisely controls the proportion of open-ended samples in the evaluation set. By adjusting α, we can assess model performance under both familiar, predefined contexts and diverse, open-ended scenarios. This dual exposure is crucial for detecting overfitting—models that have overfitted to narrow storylines will perform poorly when facing the broader diversity in the evaluation set, enabling more accurate measurement of their generalization across various narrative settings.

4.2 Discrete Diffusion Language Model Pipeline for Dataset Generation

Story visualization requires continuous sequences of 4–5 narratives, making the direct use of T_{mixed} unsuitable for evaluating narrative continuity. To address this, we propose a dataset generation framework that leverages a high-diversity Discrete Diffusion Language Model to produce coherent story sequences.

First, we construct the training dataset by combining open-ended and closed-ended data at the mixing ratio α, resulting in T_{mix}. This dataset is then processed through a tokenizer and text encoder to generate token sequences $\mathbf{z} \in R^L$ and embeddings $\mathbf{h} \in R^{L \times d_e}$, where L is the sequence length and d_e the embedding dimensionality.

We then apply a pre-trained masked language model for discrete diffusion training following the DiffusionEAGS methodology (Koh et al., 2024a). In this process, tokens are selectively masked in descending order of entropy—computed for each token embedding $\mathbf{h}_i$ as

$$\mathcal{H}(\mathbf{h}_i) = -\sum_k p_k \log p_k,$$

where p_k denotes the probability over token predictions at position i. This entropy-driven masking

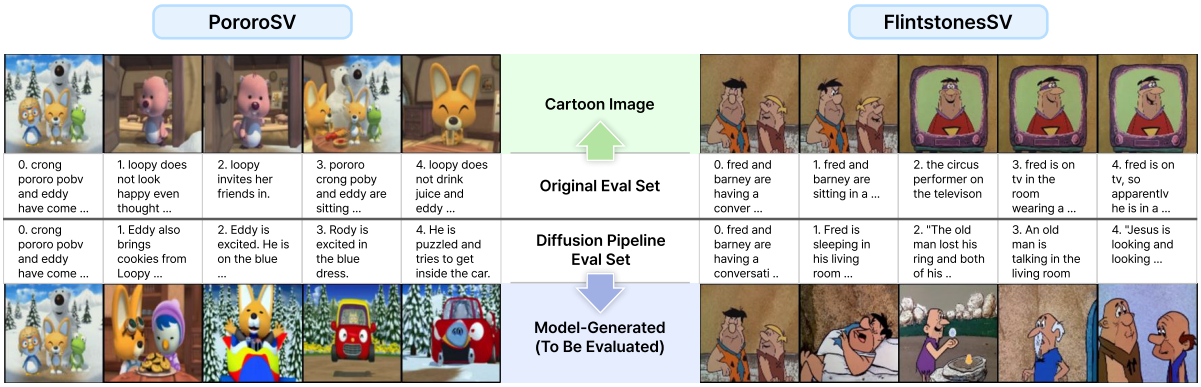

Figure 3: Sample outputs from the diffusion-augmented dataset generation pipeline for PororoSV and FlintstonesSV. The original evaluation set is enriched with open-ended elements, producing a more diverse test dataset for objective evaluation. The story visualization model utilized to generate images is AR-LDM(Pan et al., 2022).

encourages the model to generate highly variable elements, thereby enhancing output diversity. To reinforce contextual continuity and ensure compatibility with the closed-ended narratives, we incorporate character names from $\mathcal{C}_{\text{closed}}$ into both the model and tokenizer vocabularies. This diffusion-based training thus provides precise control over dataset composition and diversity.

After training, we supply an initial prompt—extracted from the original test set—to generate new sequences T'_{mix} that maintain narrative continuity while expanding the vocabulary, scenarios, and character details. These generated narratives are further refined using a large language model to correct minor grammatical issues. For evaluation, the refined sequences, along with the corresponding source caption and image, are used to assess story visualization models under the expanded context. Detailed prompts used for post-processing are provided in Appendix B

Figure 3 shows that the generated sequences cover a broader range of situations than the original dataset. In Table 1, we observe that as the mixing and diversity of the dataset increase, the MAUVE score decreases, while the vocabulary size increases.

5 Experiments

5.1 VCMS Metric Evaluation

Datasets We use the PororoSV (Li et al., 2019) and FlintstonesSV (Gupta et al., 2018) datasets—closed-ended benchmarks for story visualization. These datasets consist of story sequences from animated scenes and narratives, offering controlled, character-focused visual contexts. Details of datasets are in Appendix C.

Baselines We employ GPT-4o-2024-05-13 (OpenAI, 2023) as our primary LVLM for scoring generated images within the VCMS framework. For baseline comparisons, we utilize the AR-LDM model (Pan et al., 2022), which is specifically tailored for closed-ended story visualization tasks. Images are generated with AR-LDM and evaluated using a variety of metrics, allowing us to compare these diverse evaluation methods against our VCMS metric in terms of their effectiveness in assessing generated image quality.

Experimental Details After training each baseline on each dataset, we generate test outputs. To enhance character-specific context, tokens for character names are added to the model's vocabulary.

Evaluation We compare traditional metrics (FID, CLIP-I, CLIP-T, and Inception-v3 F1) with VCMS, and employ several LVLM-based scores: Direct LVLM Score (LVLM scores without VCMS), Multisample LVLM Score (the average of multiple scores), and Log-Prob LVLM Score (evaluation based on log probabilities of binary responses).

Correlations between human responses and metric results are computed using Spearman's ρ, Pearson's r, and Kendall's τ, with Fisher's z-transform and z-score normalization applied. More details on the human evaluation are provided in Appendix D.

5.2 Pipeline-Generated Test Dataset Evaluation

Datasets We use the original test sets of PororoSV and FlintstonesSV as control groups. To create a mixed dataset (T_{mix}), we augment these control groups with data from the StorySalon dataset (Liu et al., 2024), which contains story sequences sourced from YouTube, storybooks, and other di-

PororoSV			
Metric	$\rho \uparrow$	$\tau \uparrow$	$r \uparrow$
Human-Human	**0.4118**	**0.27**	**0.41**
FID	-0.0110	-0.008	0.002
CLIP-I	0.2448	0.16	0.22
CLIP-T	0.1307	0.08	0.14
Inception-v3	0.1104	0.08	0.12
LVLM-based score			
Direct LVLM score (single)	-0.0100	-0.01	0
Direct LVLM score (context avg)	0.0300	0.02	0.04
MultiSample LVLM score	0.1511	0.10	0.16
LogProb LVLM score	0.2877	0.19	0.26
VCMS $_{\text{0-shot}}$	0.3541	0.23	0.36
VCMS $_{\text{1-shot}}$	0.3769	0.25	0.39
VCMS $_{\text{2-shot}}$	**0.4001**	**0.27**	**0.42**
FlintstonesSV			
Metric	$\rho \uparrow$	$\tau \uparrow$	$r \uparrow$
Human-Human	**0.5361**	**0.36**	**0.51**
FID	-0.0300	-0.020	0.042
CLIP-I	0.1206	0.08	0.16
CLIP-T	0.1923	0.13	0.20
LVLM-based score			
Direct LVLM score (single)	0.1717	0.13	0.16
Direct LVLM score (context avg)	0.0601	0.04	0.08
MultiSample LVLM score	0.2877	0.20	0.31
LogProb LVLM score	0.1511	0.03	0.05
VCMS $_{\text{0-shot}}$	0.4973	0.32	0.51
VCMS $_{\text{1-shot}}$	0.4973	0.32	0.50
VCMS $_{\text{2-shot}}$	**0.5101**	**0.33**	**0.52**

Table 2: Human-Metric Correlations across PororoSV and FlintstonesSV evaluation sets (with Fisher's Z-transformed Spearman and Pearson values). *Human-Human* represents the correlation observed between evaluations conducted by different human annotators.

verse media. For experimentation, we mix the open-ended data with the control groups at varying open-ended mix ratios (α) of 25%, 50%, and 75%.

Baselines Our diffusion language model is based on the discrete diffusion approach of DiffusionEAGS (Koh et al., 2024a), with training and sampling processes adjusted accordingly. For image generation, we employ both the AR-LDM model and a finetuned Stable Diffusion Model (SDM) (Rombach et al., 2022), with each model trained on the respective datasets.

Experimental Details The training and generation process involves 8 diffusion steps, utilizing 1 A100 GPU with a batch size 64. We adopt a top-k sampling with $k = 10$ and leverage GPT-4o for post-processing to correct grammatical errors. The number of generated evaluation samples is identical to that of the original evaluation set.

PororoSV			
mix_ratio	PPL $\downarrow$	SOME $\uparrow$	TIGER $\uparrow$
Original	278.257	0.667	-6.114
mix_25	132.720	0.849	-6.525
mix_50	177.929	0.833	-6.274
mix_75	172.656	0.865	-6.709
FlintstonesSV			
Original	89.102	0.760	-7.124
mix_25	62.735	0.882	-7.297
mix_50	71.231	0.880	-7.553
mix_75	63.864	0.875	-8.061

Table 3: Text quality of pipeline-generated storylines for PororoSV and FlintstonesSV at different mixing ratios (α) compared with the original evaluation set. PPL measures perplexity, SOME indicates grammatical structure, and TIGER reflects story coherence and continuity. The reported variances of the TIGER score are 0.047 for PororoSV and 0.151 for FlintstonesSV.

Evaluation We assess generated storylines using metrics including Perplexity (GPT-2 Large/XL), SOME(Yoshimura et al., 2020), a corpus-based grammar metric, and TIGERscore(Jiang et al., 2024) for narrative coherence and continuity. We then apply our VCMS metric on both the pipeline-generated datasets—with open-ended mix ratios of 25%, 50%, and 75%—as well as the original test set. This evaluation shows while overall quality remains consistent, models overfitted to closed-ended data struggle with open-ended scenarios.

6 Results and Analysis

6.1 VCMS Metric Evaluation

Our evaluation in Table 2 shows that the VCMS metric achieves over 95% of the human–human Spearman's ρ correlation, demonstrating the effectiveness and robustness of our framework.

We assessed outputs generated by the AR-LDM (Pan et al., 2022) model using a variety of metrics. Notably, our VCMS metric consistently attains the highest human correlation—closely matching the level of agreement observed between human annotators. In addition to traditional scores, we compared several LVLM-derived metrics (including the Direct LVLM score, multisample LVLM score, and log-prob LVLM score), and found that VCMS exhibited the strongest alignment with human assessments. This indicates that its high performance is not solely due to the underlying power of the LVLM, but also to the effectiveness of the

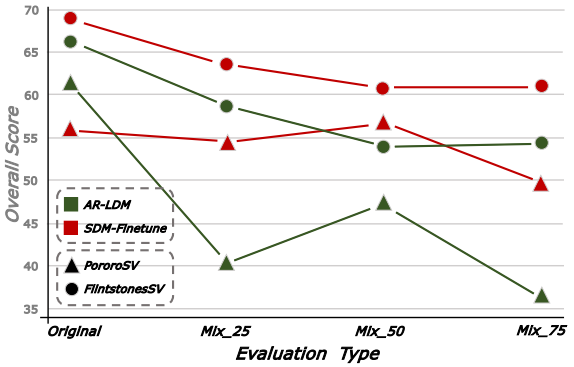

Figure 4: Overall VCMS scores at different mixing ratios for PororoSV and FlintstonesSV. AR-LDM exhibits greater sensitivity to increased open-ended content, while fine-tuned SDM maintains more consistent performance across diverse scenarios.

VCMS structure. Furthermore, our evaluation of dataset gold labels (detailed in Appendix E) reveals consistently high values across the score distribution, further supporting the reliability of the VCMS framework as a human-aligned evaluation metric for story visualization.

6.2 Pipeline-Generated Test Dataset Evaluation

Our evaluation of pipeline-generated datasets indicates that while overall narrative quality is maintained, models models overfitted to closed-ended data struggle with diverse augmented content.

6.2.1 Story Generation Quality

Our analysis of textual quality of the pipeline-generated eval set (Table 3) reveals that the generated storylines retain high quality even as open-ended mixing increases. The low perplexity indicates fluent text generation, while the SOME score confirms strong grammatical structure in the narratives produced by our diffusion model. The TIGER score, which reflects story coherence and task specificity, shows only a minor decline—expected given the increased narrative diversity and broader vocabulary introduced by open-ended mixing. Importantly, the MAUVE score (Table 1) decreases with higher open-ended mixing ratios, demonstrating that our pipeline successfully produces high-quality storylines that integrate a broader range of vocabulary, situations, and characters.

6.2.2 Evaluation on Generated Dataset

Figure 4 illustrates that overall evaluation scores drop as the proportion of open-ended content increases. AR-LDM's scores vary significantly

Dataset	PororoSV			FlintstonesSV		
	$\rho\uparrow$	$\tau\uparrow$	$r\uparrow$	$\rho\uparrow$	$\tau\uparrow$	$r\uparrow$
GPT-4o *0-shot*	0.3530	0.23	0.36	0.4973	0.32	0.51
GPT-4o *1-shot*	0.3769	0.25	0.39	0.4973	0.32	0.50
GPT-4o *2-shot*	**0.4001**	**0.27**	**0.42**	**0.5101**	**0.33**	**0.52**
GPT-4o *3-shot*	0.1968	0.14	0.23	0.2960	0.21	0.32
Gemini-Pro *0-shot*	0.3530	0.24	0.34	0.3884	0.26	0.39
Gemini-Pro *1-shot*	0.1811	0.13	0.18	0.2350	0.16	0.24
LLAVA *0-shot*	0.0601	0.05	0.09	0.1307	0.09	-0.004
LLAVA *1-shot*	0.0500	0.04	0.05	0.1717	0.13	0.19

Table 4: Human correlation of VCMS score generated by diverse LVLMs (with Fisher's Z-transformed Spearman and Pearson values).

across mix ratios, reflecting sensitivity to open-ended data, while the SDM-based model (Rombach et al., 2022) maintains consistent quality across diverse scenarios. This suggests that narrative continuity affects the generated image quality of the model differently. On PororoSV, a slight score boost at $\alpha = 0.5$ was reported; this may be linked to a higher MAUVE score in Table 1.

To assess how image style, continuity, and overall performance depend on textual storylines, it is crucial to evaluate models using a pipeline like ours. This shows that those overfitted to closed-ended narratives generalize poorly to open-ended scenarios. Full VCMS scores are in Appendix F.

6.3 Various LVLMs

As shown in Table 4, GPT-4o (OpenAI, 2023) shows performance gains up to 3-shot prompting, whereas Gemini-1.5-Pro (Team et al., 2024) performs comparably in a 0-shot setting but declines with 1-shot prompts. In contrast, LLAVA (Liu et al., 2023)[1] consistently underperforms regardless of the number of few-shot examples. Interestingly, all evaluated models yield higher scores on FlintstonesSV compared to PororoSV, suggesting that differences in dataset representation—possibly due to greater exposure to FlintstonesSV in Western media—affect the evaluation outcomes.

6.4 Comparing Models via New Eval Methods

Our analysis reveals that while models trained on fixed narratives excel in closed-ended settings, they falter with open-ended data, highlighting the need to evaluate both overfitting and generalization.

[1]We downloaded the pretrained weights of LLAVA from https://huggingface.co/llava-hf/llava-1.5-7b-hf

Method	PororoSV		FlintstonesSV	
	Original	Augmented	Original	Augmented
Training Set	74.46		78.60	
SDM	22.76	36.15	40.57	39.43
prompt-SDM	24.16	37.91	43.36	41.23
Finetune-SDM	59.31	56.11	68.47	60.24
Story-LDM	26.44	17.88	–	–
StoryDall-E	31.74	28.09	29.17	26.68
AR-LDM	60.87	46.60	65.79	53.31

Table 5: Combined overall VCMS scores for various baselines on the original and diffusion-augmented evaluation sets (mix ratio $\alpha = 0.5$). The "Training Set" row indicates the gold label score for each training set.

Table 5 summarizes the combined VCMS scores for various story visualization models on the PororoSV and FlintstonesSV datasets, evaluated on both the original and diffusion-augmented sets. On the original set, both AR-LDM (main) and fine-tuned SDM achieve competitive scores, reflecting strong performance in handling structured narratives. In contrast, Story-LDM (Rahman et al., 2023), StoryDall-E (Maharana et al., 2022), and SDM without fine-tuning exhibit relatively lower scores, indicating limitations in context reflection and visual continuity.

On the diffusion-augmented set, which introduces more diverse narratives, fine-tuned SDM shows a notable adaptability, while AR-LDM's performance declines. This contrast suggests that models heavily trained on fixed storylines excel in controlled settings but struggle with broader, open-ended scenarios. Notably, AR-LDM is highly specialized due to its training on a distinct dataset, whereas SDM—as a pre-trained model—performs strongly with additional fine-tuning on diffusion-augmented data; indeed, on PororoSV, the performance of SDM without fine-tuning even improves with open-ended content.

Overall, these findings highlight each model's strengths and weaknesses, underscoring the need to evaluate performance under both structured and diverse narrative conditions. Detailed VCMS breakdowns for each aspect are provided in Appendix G.

7 Related Works

7.1 Diffusion Language Models

Diffusion models have become influential since DDPM (Ho et al., 2020). In natural language processing, approaches include both continuous and discrete diffusion models. Early work (Li et al., 2022b; Gong et al., 2023a,b) mapped tokenized sequences into embedding spaces using pre-trained classifiers. However, recent discrete diffusion models—such as D3PM (Austin et al., 2023), DiffusionBERT (He et al., 2022), and SEDD (Lou et al., 2024)—address the discrete nature of language. Furthermore, recent studies (Shi et al., 2024; Zheng et al., 2024; Sahoo et al., 2024) streamlined these methods, effectively bridging the gap between discrete diffusion and masked language models.

7.2 LVLM Scoring

Traditional metrics like Inception Score (Salimans et al., 2016) and FID (Heusel et al., 2017) assess image quality, not text–image alignment. To address this, metrics such as ClipScore (Hessel et al., 2021) and BlipScore (Li et al., 2022a) have been developed, with further enhancements incorporating Mutual Information Divergence (Kim et al., 2022) and R-Precision (Park et al., 2021), as well as holistic measures (Cho et al., 2023; Lee et al., 2024). Recently, the rise of multimodal large language models like GPT-4 (Zhang et al., 2023) has spurred new evaluation methods (Lu et al., 2024; Ku et al., 2023; Peng et al., 2024). While techniques such as Chain-of-Thought (Huang et al., 2023) and fine-tuning for consistency (Wu et al.) have improved prompt fidelity, few approaches address narrative coherence across story sets. Inspired by work on video continuity (Krojer et al., 2024), we propose VCMS to jointly assess text-image alignment and inter-image coherence in story visualization.

8 Conclusions

In this paper, we introduced the Visual Context-Aware Metric for Story Visualization (VCMS) and a diffusion-augmented evaluation pipeline that together provide a human-aligned, robust framework for assessing both text-image alignment and inter-image coherence. Our experiments show that VCMS achieves over 95% correlation with human judgments, outperforming traditional and LVLM-derived metrics, and reveal that while models like AR-LDM excel in structured settings, pre-trained models such as SDM demonstrate superior adaptability to diverse, open-ended narratives. Overall, our work offers a scalable evaluation approach that highlights each model's strengths and limitations, paving the way for future advancements in generating coherent visual stories.

Limitations

Our work has several limitations. First, the VCMS metric relies on GPT-4o and other LVLMs, which may inherit biases and have limited interpretability, potentially affecting evaluation outcomes. Second, our experiments primarily use closed-ended datasets augmented with open-ended data at controlled mixing ratios; while this offers a degree of narrative diversity, it may not fully capture the richness of real-world storytelling. Finally, the scalability and generalizability of our framework require further validation on additional datasets and in varied application contexts.

Ethical Statements

Our research involves large-scale datasets that may contain inherent biases, which can be inadvertently reinforced by automated evaluation frameworks. The reliance on pre-trained models such as GPT-4o raises concerns regarding model bias, fairness, and transparency in evaluation outcomes. We advocate for ongoing efforts to mitigate these biases, including the use of diverse datasets and the implementation of robust bias detection and mitigation strategies, to ensure that the deployment of our evaluation framework in story visualization research is both ethical and responsible. We confirm that all participants for human evaluation were informed about the study's purpose and provided consent. The data collection protocol was compliant with institutional guidelines, and all annotators were fairly compensated for their labor. AI assistants were used to support language polishing and limited coding assistance. They were not used to generate the core research ideas, design the experiments, make methodological decisions, interpret the results.

Acknowledgments

This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea goverment(MSIT) [RS-2025-02263628]. This work was partly supported by Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) [RS-2021-II211343, Artificial Intelligence Graduate School Program (Seoul National University) & RS-2021-II212068, Artificial Intelligence Innovation Hub (Artificial Intelligence Institute, Seoul National University)], K. Jung is with ASRI, Seoul National University, Korea.

References

Jacob Austin, Daniel D. Johnson, Jonathan Ho, Daniel Tarlow, and Rianne van den Berg. 2023. Structured denoising diffusion models in discrete state-spaces. *Preprint*, arXiv:2107.03006.

Jiaao Chen and Diyi Yang. 2023. Controllable conversation generation with conversation structures via diffusion models. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 7238–7251, Toronto, Canada. Association for Computational Linguistics.

Xiuyuan Chen, Yuan Lin, Yuchen Zhang, and Weiran Huang. 2024. Autoeval-video: An automatic benchmark for assessing large vision language models in open-ended video question answering. In *European Conference on Computer Vision*, pages 179–195. Springer.

Jaemin Cho, Yushi Hu, Roopal Garg, Peter Anderson, Ranjay Krishna, Jason Baldridge, Mohit Bansal, Jordi Pont-Tuset, and Su Wang. 2023. Davidsonian scene graph: Improving reliability in fine-grained evaluation for text-image generation. *arXiv preprint arXiv:2310.18235*.

Shangbin Feng, Chan Young Park, Yuhan Liu, and Yulia Tsvetkov. 2023. From pretraining data to language models to downstream tasks: Tracking the trails of political biases leading to unfair NLP models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11737–11762, Toronto, Canada. Association for Computational Linguistics.

Ronald A Fisher. 1915. Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. *Biometrika*, 10(4):507–521.

Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. 2023. Gptscore: Evaluate as you desire. *arXiv preprint arXiv:2302.04166*.

Shansan Gong, Mukai Li, Jiangtao Feng, Zhiyong Wu, and Lingpeng Kong. 2023a. Diffuseq: Sequence to sequence text generation with diffusion models. *Preprint*, arXiv:2210.08933.

Shansan Gong, Mukai Li, Jiangtao Feng, Zhiyong Wu, and Lingpeng Kong. 2023b. Diffuseq-v2: Bridging discrete and continuous text spaces for accelerated seq2seq diffusion models. *Preprint*, arXiv:2310.05793.

Tanmay Gupta, Dustin Schwenk, Ali Farhadi, Derek Hoiem, and Aniruddha Kembhavi. 2018. Imagine this! scripts to compositions to videos. *Preprint*, arXiv:1804.03608.

Zhengfu He, Tianxiang Sun, Kuanning Wang, Xuanjing Huang, and Xipeng Qiu. 2022. Diffusionbert: Improving generative masked language models with diffusion models. *Preprint*, arXiv:2211.15029.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. 2021. Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718*.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2018. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Preprint*, arXiv:1706.08500.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. *Preprint*, arXiv:2006.11239.

Kaiyi Huang, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. 2023. T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation. *Advances in Neural Information Processing Systems*, 36:78723–78747.

Dongfu Jiang, Yishan Li, Ge Zhang, Wenhao Huang, Bill Yuchen Lin, and Wenhu Chen. 2024. Tigerscore: Towards building explainable metric for all text generation tasks. *Preprint*, arXiv:2310.00752.

Jin-Hwa Kim, Yunji Kim, Jiyoung Lee, Kang Min Yoo, and Sang-Woo Lee. 2022. Mutual information divergence: A unified metric for multimodal generative models. *Advances in Neural Information Processing Systems*, 35:35072–35086.

Hyukhun Koh, Minha Jhang, Dohyung Kim, Sangmook Lee, and Kyomin Jung. 2024a. Plm-based discrete diffusion language models with entropy-adaptive gibbs sampling. *Preprint*, arXiv:2411.06438.

Hyukhun Koh, Dohyung Kim, Minwoo Lee, and Kyomin Jung. 2024b. Can LLMs recognize toxicity? a structured investigation framework and toxicity metric. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 6092–6114, Miami, Florida, USA. Association for Computational Linguistics.

Benno Krojer, Dheeraj Vattikonda, Luis Lara, Varun Jampani, Eva Portelance, Christopher Pal, and Siva Reddy. 2024. Learning action and reasoning-centric image editing from videos and simulations. *arXiv preprint arXiv:2407.03471*.

Max Ku, Dongfu Jiang, Cong Wei, Xiang Yue, and Wenhu Chen. 2023. Viescore: Towards explainable metrics for conditional image synthesis evaluation. *arXiv preprint arXiv:2312.14867*.

Max Ku, Dongfu Jiang, Cong Wei, Xiang Yue, and Wenhu Chen. 2024. Viescore: Towards explainable metrics for conditional image synthesis evaluation. *Preprint*, arXiv:2312.14867.

Tony Lee, Michihiro Yasunaga, Chenlin Meng, Yifan Mai, Joon Sung Park, Agrim Gupta, Yunzhi Zhang, Deepak Narayanan, Hannah Teufel, Marco Bellagente, et al. 2024. Holistic evaluation of text-to-image models. *Advances in Neural Information Processing Systems*, 36.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022a. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR.

Xiang Lisa Li, John Thickstun, Ishaan Gulrajani, Percy Liang, and Tatsunori B. Hashimoto. 2022b. Diffusion-lm improves controllable text generation. *Preprint*, arXiv:2205.14217.

Yitong Li, Zhe Gan, Yelong Shen, Jingjing Liu, Yu Cheng, Yuexin Wu, Lawrence Carin, David Carlson, and Jianfeng Gao. 2019. Storygan: A sequential conditional gan for story visualization. *Preprint*, arXiv:1812.02784.

Zhiqiu Lin, Deepak Pathak, Baiqi Li, Jiayao Li, Xide Xia, Graham Neubig, Pengchuan Zhang, and Deva Ramanan. 2024. Evaluating text-to-visual generation with image-to-text generation. In *European Conference on Computer Vision*, pages 366–384. Springer.

Chang Liu, Haoning Wu, Yujie Zhong, Xiaoyun Zhang, Yanfeng Wang, and Weidi Xie. 2024. Intelligent grimm – open-ended visual storytelling via latent diffusion models. *Preprint*, arXiv:2306.00973.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *Preprint*, arXiv:2304.08485.

Aaron Lou, Chenlin Meng, and Stefano Ermon. 2024. Discrete diffusion modeling by estimating the ratios of the data distribution. *Preprint*, arXiv:2310.16834.

Yujie Lu, Xianjun Yang, Xiujun Li, Xin Eric Wang, and William Yang Wang. 2024. Llmscore: Unveiling the power of large language models in text-to-image synthesis evaluation. *Advances in Neural Information Processing Systems*, 36.

Adyasha Maharana, Darryl Hannan, and Mohit Bansal. 2021. Improving generation and evaluation of visual stories via semantic consistency. *Preprint*, arXiv:2105.10026.

Adyasha Maharana, Darryl Hannan, and Mohit Bansal. 2022. Storydall-e: Adapting pretrained text-to-image transformers for story continuation. *Preprint*, arXiv:2209.06192.

OpenAI. 2023. GPT-4 technical report. *CoRR*, abs/2303.08774.

Xichen Pan, Pengda Qin, Yuhong Li, Hui Xue, and Wenhu Chen. 2022. Synthesizing coherent story with auto-regressive latent diffusion models. *Preprint*, arXiv:2211.10950.

Dong Huk Park, Samaneh Azadi, Xihui Liu, Trevor Darrell, and Anna Rohrbach. 2021. Benchmark for compositional text-to-image synthesis. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*.

Yuang Peng, Yuxin Cui, Haomiao Tang, Zekun Qi, Runpei Dong, Jing Bai, Chunrui Han, Zheng Ge, Xiangyu Zhang, and Shu-Tao Xia. 2024. Dreambench++: A human-aligned benchmark for personalized image generation. *arXiv preprint arXiv:2406.16855*.

Krishna Pillutla, Lang Liu, John Thickstun, Sean Welleck, Swabha Swayamdipta, Rowan Zellers, Sewoong Oh, Yejin Choi, and Zaid Harchaoui. 2023. Mauve scores for generative models: Theory and practice. *Preprint*, arXiv:2212.14578.

Tanzila Rahman, Hsin-Ying Lee, Jian Ren, Sergey Tulyakov, Shweta Mahajan, and Leonid Sigal. 2023. Make-a-story: Visual memory conditioned consistent story generation. *Preprint*, arXiv:2211.13319.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. *Preprint*, arXiv:2112.10752.

Subham Sekhar Sahoo, Marianne Arriola, Yair Schiff, Aaron Gokaslan, Edgar Marroquin, Justin T Chiu, Alexander Rush, and Volodymyr Kuleshov. 2024. Simple and effective masked diffusion language models. *arXiv preprint arXiv:2406.07524*.

Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. 2016. Improved techniques for training gans. *Advances in neural information processing systems*, 29.

Jiaxin Shi, Kehang Han, Zhe Wang, Arnaud Doucet, and Michalis K Titsias. 2024. Simplified and generalized masked diffusion for discrete data. *arXiv preprint arXiv:2406.04329*.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. 2015. Rethinking the inception architecture for computer vision. *Preprint*, arXiv:1512.00567.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

Haoning Wu, Xiele Wu, Chunyi Li, Zicheng Zhang, Chaofeng Chen, Xiaohong Liu, Guangtao Zhai, and Weisi Lin. T2i-scorer: Quantitative evaluation on text-to-image generation via fine-tuned large multimodal models. In *ACM Multimedia 2024*.

Dingyi Yang and Qin Jin. 2024. What makes a good story and how can we measure it? a comprehensive survey of story evaluation. *Preprint*, arXiv:2408.14622.

Ryoma Yoshimura, Masahiro Kaneko, Tomoyuki Kajiwara, and Mamoru Komachi. 2020. SOME: Reference-less sub-metrics optimized for manual evaluations of grammatical error correction. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6516–6522, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Xinlu Zhang, Yujie Lu, Weizhi Wang, An Yan, Jun Yan, Lianke Qin, Heng Wang, Xifeng Yan, William Yang Wang, and Linda Ruth Petzold. 2023. Gpt-4v (ision) as a generalist evaluator for vision-language tasks. *arXiv preprint arXiv:2311.01361*.

Kaiwen Zheng, Yongxin Chen, Hanzi Mao, Ming-Yu Liu, Jun Zhu, and Qinsheng Zhang. 2024. Masked diffusion models are secretly time-agnostic masked models and exploit inaccurate categorical sampling. *arXiv preprint arXiv:2409.02908*.

A Example Prompts for VCMS Metric

In this section, we provide example prompts used to evaluate model outputs according to the VCMS metric, which focuses on Caption Reflection and Image Consistency. Figure 5 shows the prompts in detail, covering different evaluation criteria such as emotions/behavior, setting, consistency, and detailing. These prompts serve as standardized queries posed to GPT-4o to score each generated image's alignment with the caption and continuity between scenes. Each prompt asks specific questions to ensure a thorough assessment of visual storytelling aspects.

B Example Prompts for Post-Processing Test Set Generation

This section includes prompts used for post-processing the test set. Figure 6 provides an example prompt used to correct grammar in generated captions. In this process, GPT-4o is directed to output only the corrected sentence, ensuring grammatical accuracy in test set captions, which helps maintain consistency and clarity in model evaluations.

C PororoSV and FlintstonesSV

This section provides an overview of the **PororoSV** and **FlintstonesSV** datasets, two popular benchmarks for cartoon-based story visualization:

- **PororoSV**:
  - *Number of Stories (samples)*: 15,336
  - *Images per Story (avg.)*: 6.92
  - *Domain*: Cartoon ("Pororo the Little Penguin")
- **FlintstonesSV**:
  - *Number of Stories (samples)*: 45,212
  - *Images per Story (avg.)*: 8.31
  - *Domain*: Cartoon ("The Flintstones")

Both datasets feature continuous storylines presented through sequential images, making them well-suited for evaluating narrative coherence and visual consistency in story visualization tasks.

D Detail of human evaluation

We recruited graduate students fluent in English through the university's community. The recruited

```
1. Emotions/Behavior: On a scale from 0 to 1, how accurately does the character's emotions and behavior in the generated image reflect the emotions and behavior described in the caption? Pay attention to facial expressions, body language, and actions. And consider whether direct behaviors have been reflected.

2. Setting: On a scale from 0 to 1, how appropriate is the background setting in the generated image compared to what is described in the caption?

3. Consistency: On a scale from 0 to 1, how well does the appearance of the character in the generated image match the appearance of the same character in the previous scene? Consider factors like clothing, facial features, and overall design. If the character doesn't match, consider styles of drawings. Pay attention to facial expressions, body language, and actions.

4. Detailing: On a scale from 0 to 1, how well does the background in the generated image maintain consistent artistic style compared to the previous scene? Consider the level of detail, color schemes, and any recurring elements.


Suffix : Do not add any explains, comments, or suggestions. Just answer the result by a real number.
```

Figure 5: Example prompts for VCMS.

```
{"role": "system", "content": "You are a helpful assistant that corrects grammar."},
{"role": "user", "content": f"Correct the following sentence: text. only give corrected sentence as an output"}
```

Figure 6: Example prompts for post-processing test set generation.

annotators were provided with a detailed description of task definitions, instructions, and samples of each model. Also, all applicants were informed that their annotations would be used for academic purposes and would be published in paper material through the recruitment announcement and instructions.
A total of 7 annotators were provided with 2000 full samples including source and generated images and their captions. You can find the samples used for the survey at https://www.2024-cvpr.p-e.kr/. For the payment of the annotators, the co-authors conducted annotations for 5 hours first to estimate the average number of annotations that could be completed in the same time. Based on this estimation, a rate of 17.88 dollars per hour was established to ensure that the annotators would be paid at least the minimum wage.

We conduct human evaluation on different ranges of samples for different annotators. As each annotator has their own subjective perspective for each metric, their scores exhibit different distributions. For example, one annotator may respond with a score of '1' for the 'Caption Reflection' question when viewing an image of Loopy standing in front of a door, accompanied by the caption "Loopy invites her friends in," as they feel the image provides sufficient evidence to reflect the caption. However, another annotator may respond with '0.5' for the same metric and sample, as the image does not depict Loopy's friends and thus only partially reflects the caption. Therefore, we apply z-normalization within each annotator's set of scores to account for individual scoring tendencies.

Additionally, we apply Fisher's Z-Transform (Fisher, 1915) to each Pearson and Spearman correlation result to obtain more normalized outcomes.

E Gold Label Evaluation with VCMS

Table 6 presents VCMS scores computed on the gold labels of the test set, compared with outputs generated by AR-LDM. The scores are broken down into several key aspects: character context, background context, emotion & behavior, and overall performance. These results demonstrate that the gold labels consistently achieve higher VCMS scores than AR-LDM outputs across all evaluated dimensions, highlighting the effectiveness of VCMS in capturing detailed narrative and visual continuity aspects.

F VCMS Score of AR-LDM and SDM-Finetune for Diffusion-Augmented Eval Set

In Table 7 and 8, we present VCMS scores across various criteria (Character Context, Background Context, Emotion & Behavior, Background, Overall) for the AR-LDM model and SDM-finetuned model on the Diffusion-Augmented Evaluation Set. Scores are reported for PororoSV and FlintstonesSV benchmarks at different mix levels (mix_25, mix_50, mix_75), illustrating how model performance adapts to increasing narrative diversity.

G Detailed results for Comparing Models via New Eval Methods

In Table 9 and 10, we present VCMS scores across various criteria (Character Context, Background Context, Emotion & Behavior, Background, Overall) for the diverse story visualizatoin model on both original and Diffusion-Augmented Evaluation Set. Scores are reported for PororoSV and FlintstonesSV benchmarks.

H Human Correlation of VCMS score for specific sections

This section presents human correlation data for specific sections of the VCMS evaluation. Table 11 summarizes the correlation results, indicating how closely VCMS scores align with human evaluations in different aspects of story visualization, such as character consistency, background quality, and overall narrative coherence. These correlations provide insight into the effectiveness of VCMS as a reliable metric for capturing the nuances in human judgment.

I LLM augmented dataset

To further contextualize the contribution of our proposed diffusion-based augmentation pipeline, we conducted a small-scale comparison with a large language model (LLM)-based augmentation method. Specifically, we used GPT-4o to generate alternative captions for a subset of the *PororoSV* dataset, constructing a toy augmentation baseline to assess model robustness under LLM-induced variation.

We evaluated two representative models—**AR-LDM** and **finetuned-SDM**—on this GPT-augmented dataset. The results are summarized

	Char. Context		Back. Context		Emotion&Behavior		Background		Overall	
Dataset	gold label	AR-LDM	gold label	AR-LDM	gold label	AR-ADM	gold label	AR-LDM	gold label	AR-LDM
PororoSV	69.76	56.53	86.56	68.06	54.62	42.15	86.88	76.72	74.46	60.87
FlintstonesSV	66.13	46.91	88.37	71.47	76.03	66.48	83.88	78.28	78.60	65.79

Table 6: VCMS scores calculated on the gold labels of the test set compared with outputs generated by AR-LDM.

Dataset	**PororoSV**					**FlintstonesSV**				
	Char. Context	**Back. Context**	**Emotion & Behavior**	**Background**	**Overall**	**Char. Context**	**Back. Context**	**Emotion & Behavior**	**Background**	**Overall**
Original ($\alpha = 0$)	56.54	68.07	42.15	76.73	60.86	46.91	71.47	66.49	78.29	65.79
				Diffusion-Augmented Eval Set						
$\alpha = 0.25$	24.41	53.01	19.99	10.45	39.84	24.67	64.92	64.62	78.66	58.22
$\alpha = 0.5$	31.80	53.66	31.74	69.44	46.60	19.95	61.57	60.11	71.61	53.31
$\alpha = 0.75$	17.08	41.62	22.76	61.27	35.68	18.86	50.07	57.53	76.20	53.71

Table 7: VCMS score of AR-LDM Model for Diffusion-Augmented Test Dataet

in Table 12.

These results suggest that while both models are capable of handling LLM-generated variation to some extent, AR-LDM demonstrates stronger robustness under such conditions.

However, the primary strength of our diffusion-based augmentation approach lies not in outperforming LLM-generated baselines, but in offering **systematic controllability** over narrative complexity. As detailed in Figure 4 of the main text, our method produces evaluation sets with varying levels of narrative mixture—*Mix_25*, *Mix_50*, and *Mix_75*—which allow for fine-grained analysis of model generalization under increasingly open-ended and compositional conditions.

We observed clear and interpretable performance trends across mixture levels, particularly a **gradual performance decline** for AR-LDM as narrative complexity increased. This supports our claim that discrete diffusion models enable principled and scalable *stress-testing* of story visualization systems, going beyond what LLM-based augmentation alone can offer.

This analysis highlights the complementary nature of diffusion models as an augmentation tool and justifies their use for creating challenging, diverse, and controllable evaluation benchmarks in the story visualization domain.

J Generalization to Non-Character / Non-Cartoon Narratives and Q1/Q3 Bias

To further validate the general applicability of VCMS beyond character-centric cartoon narratives, we conducted two toy-set experiments using the Short-Films 20K (SF20K) dataset. These experiments are designed to test whether VCMS can robustly separate caption fidelity from inter-image consistency under realistic, non-cartoon, and non-character-heavy settings.

Specifically, we compare a normal evaluation set with a 50% corrupted set, where captions are randomly corrupted for half of the samples. If VCMS correctly disentangles caption fidelity from visual consistency, the fidelity-related components should decrease substantially under caption corruption, while the consistency-related components should remain relatively stable.

J.1 Non-Cartoon Generalization Test

We first evaluate VCMS on realistic short-film data to verify whether the metric generalizes beyond cartoon-style story visualization datasets. Table 13 reports the mean VCMS component scores on the normal and 50% corrupted sets.

The results show a clear separation between caption fidelity and image consistency. The fidelity-related components, Q1 and Q2, drop sharply by 48.8% and 46.2%, respectively, when captions are corrupted. This confirms that VCMS accurately captures caption-image alignment even in realistic, non-cartoon environments.

In contrast, the consistency-related components, Q3 and Q4, remain highly stable, changing by only -0.7% and $+3.4\%$, respectively. Since caption corruption should not substantially alter the visual continuity between consecutive frames, this stability indicates that VCMS measures inter-image consistency independently from caption fidelity. These results support the general applicability of VCMS beyond animated or cartoon-based datasets.

J.2 Non-Character Bias Test

We further examine whether VCMS introduces spurious bias in scenes where explicit characters are absent. This is important because Q1 and Q3 are originally phrased around character behavior and

Dataset	PororoSV					FlintstonesSV				
	Char. Context	Back. Context	Emotion & Behavior	Background	Overall	Char. Context	Back. Context	Emotion & Behavior	Background	Overall
Original	58.06	77.04	12.95	69.74	59.31	63.75	85.57	50.45	74.09	68.47
Diffusion-Augmented Eval Set										
mix_25	67.84	79.78	15.47	52.77	53.97	52.73	81.83	43.87	73.77	63.05
mix_50	67.30	79.33	21.10	56.72	56.11	45.04	81.30	45.02	69.60	60.24
mix_75	63.69	76.45	13.70	43.01	49.21	50.10	80.44	40.61	70.90	60.51

Table 8: VCMS score of SDM-finetune for Diffusion-Augmented Test Dataset

Model	PororoSV					FlintstonesSV				
	Char. Context	Back. Context	Emotion & Behavior	Background	Overall	Char. Context	Back. Context	Emotion & Behavior	Background	Overall
Original Data	69.76	86.56	54.62	86.88	74.46	66.13	88.37	76.03	83.88	78.60
SDM	9.67	20.1	15.9	45.4	22.76	10.94	21.09	51.1	**79.1**	40.57
prompt-SDM	9.71	18.64	17.3	50.1	24.16	15.56	31.33	51.5	75.1	43.36
Finetune-SDM	**58.07**	**77.04**	12.95	69.74	59.31	**63.75**	**85.57**	50.5	74.1	**68.47**
StoryViz	24.71	29.6	12.17	44.98	27.87	-	-	-	-	-
Story-LDM	25.73	26.98	12.38	40.67	26.44	-	-	-	-	-
StoryDall-E	23.16	45.86	8.5	49.4	31.74	27.23	52.43	7.6	29.4	29.17
AR-LDM (main)	56.53	68.06	**42.2**	**76.7**	**60.87**	46.91	71.47	**66.5**	**78.3**	65.79

Table 9: VCMS score of Various Baselines for Original Eval Set

character consistency. To test this, we construct a specialized subset of SF20K in which the main visual subject is not an explicit human or cartoon character. Table 14 shows the resulting VCMS scores.

The normal set achieves a high Q1 score of 0.927, despite the absence of explicit characters. This indicates that the LVLM does not restrict Q1 to human or cartoon-character emotion alone. Instead, when characters are absent, Q1 is naturally interpreted as the fidelity of the main object's state, action, or visually described condition in the caption.

When captions are corrupted, Q1 decreases sharply from 0.927 to 0.470, corresponding to a 49.4% drop. This confirms that Q1 remains a robust caption-fidelity measure regardless of whether the subject is a person, a cartoon character, an object, or the scene itself. Q2 similarly decreases by 43.0%, further showing that VCMS captures caption-level perturbations in non-character scenes.

Most importantly, Q3 remains stable, decreasing by only 3.2%, while Q4 changes by only $+1.1\%$. This suggests that the character-consistency question does not introduce a spurious penalty or artificial bias when explicit characters are absent. Instead, the LVLM appears to evaluate the continuity of the dominant visual subject or scene-level appearance when no character is present. Therefore, the non-character experiment supports that VCMS does not rely on a character-specific bias and can be applied to broader narrative visualization settings.

Model	PororoSV					FlintstonesSV				
	Char. Context	Back. Context	Emotion & Behavior	Background	Overall	Char. Context	Back. Context	Emotion & Behavior	Background	Overall
SDM	6.65	21.18	40.49	76.28	36.15	4.74	18.79	55.47	**78.73**	39.43
prompt-SDM	7.53	19.77	**45.54**	**78.81**	37.91	8.67	27.68	52.48	76.08	41.23
Finetune-SDM	**67.30**	**56.72**	21.10	56.72	**56.11**	**45.04**	**81.30**	45.02	69.61	**60.24**
Story-LDM	13.59	21.33	5.1	31.54	17.88	-	-	-	-	-
StoryDall-E	19.14	40.50	8.26	44.45	28.09	19.60	43.33	10.41	29.35	26.68
AR-LDM (main)	31.80	53.66	31.74	69.44	46.60	19.95	61.57	**60.11**	71.61	53.31

Table 10: VCMS score of Various Baselines for Diffusion-Augmented Eval Set

	PororoSV	FlintstonesSV
Char. Context	0.41	0.60
Back. Context	0.54	0.34
Emotion&Behavior	0.15	0.37
Background	0.12	0.24

Table 11: Human correlation of VCMS score for specific sections

Model	Mean Score
AR-LDM	0.530
Finetuned-SDM	0.468

Table 12: Model performance on GPT-4o-augmented PororoSV dataset.

VCMS Component	Normal Set	50% Corrupted Set	Change
Q1: Emotion & Behavior	0.893	0.457	↓48.8%
Q2: Setting	0.837	0.450	↓46.2%
Q3: Char. Consistency	0.691	0.686	↓0.7%
Q4: Back. Consistency	0.655	0.677	↑3.4%
Overall VCMS Score	0.793	0.571	↓28.0%

Table 13: VCMS scores on the SF20K non-cartoon generalization test. Caption corruption sharply decreases fidelity-related components, while consistency-related components remain stable.

VCMS Component	Normal Set	50% Corrupted Set	Change
Q1: Emotion & Behavior	0.927	0.470	↓49.4%
Q2: Setting	0.900	0.513	↓43.0%
Q3: Char. Consistency	0.714	0.691	↓3.2%
Q4: Back. Consistency	0.861	0.870	↑1.1%
Overall VCMS Score	0.866	0.623	↓28.2%

Table 14: VCMS scores on the SF20K non-character bias test. Even when explicit characters are absent, Q1 remains sensitive to caption corruption, while Q3 remains stable.