VISIAR: Empower MLLM for Visual Story Ideation

Zhaoyang Xia[1], Somdeb Sarkhel[2], Mehrab Tanjim[2], Stefano Petrangeli[2], Ishita Dasgupta[2], Yuxiao Chen[1], Jinxuan Xu[1], Di Liu [1], Saayan Mitra[2], Dimitris N. Metaxas[1],

[1]Rutgers University, [2]Adobe Research,

{zx149,yc984,jinxuan.xu,dl1014,dnm}@rutgers.edu, {sarkhel,tanjim,petrange,idasgupt,smitra}@adobe.com

Abstract

Ideation, the process of forming ideas from concepts, is a big part of the content creation process. However, the noble goal of helping visual content creators by suggesting meaningful sequences of visual assets from a limited collection is challenging. It requires a nuanced understanding of visual assets and the integration of open-world knowledge to support creative exploration. Despite its importance, this task has yet to be explored fully in existing literature. To fill this gap, we propose **Visual Story Ideation**, a novel and underexplored task focused on the automated selection and arrangement of visual assets into coherent sequences that convey expressive storylines.

We also present **VISIAR**, **V**isual **I**deation through **S**equence **I**ntegration and **A**sset **R**earrangement, a robust framework leveraging Multimodal Large Language Models (MLLMs), and a novel Story Graph mechanism. Our framework operates in three key stages: visual content understanding, candidate asset selection, and asset rearrangement via MLLMs. In addition, we curated a new benchmark dataset, called VTravel, to evaluate our methods both qualitatively and quantitatively. User studies and GPT-as-the-judge evaluation show that our approach surpasses GPT-4o based baseline by an average of $\mathbf{33.5}\%$ and $\mathbf{18.5}\%$ across three different metrics, demonstrating the effectiveness of our framework for generating compelling visual stories. [1]

1 Introduction

Visual story (Huang et al., 2016) plays a pivotal role across diverse domains, including digital marketing, recreational content creation, and personal content generation for social media, due to its ability to captivate audiences. However, content creators often face significant challenges in crafting expressive and coherent visual stories from their personalized assets, as the asset collection can contain several hours of content, making it overwhelming to sift through manually. More importantly, ideating compelling storylines from hundreds or thousands of assets is time-consuming and labor-intensive, requiring professional expertise in interpreting visual content, selecting relevant elements, and arranging them into an expressive storyline. Automating the ideation process can not only accelerate visual storytelling but also inspire creators, broaden access to novice users, and expand creative possibilities across industries. While research on visual storytelling (Hong et al., 2023) is of great interest, this automatic ideation process remains largely underexplored. To bridge this gap, we introduce a novel task, **Visual Story Ideation**, aimed at automatically deriving underlying visual stories from a collection of visual assets.

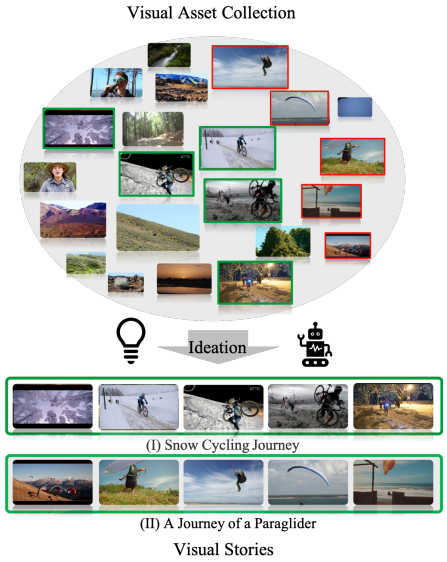


Figure 1: **Visual Story Ideation**: Given a collection of visual assets, we produce multiple potential storylines by selecting and combining visual assets from the collection automatically.

[1]Please check our project website for more details: https://github.com/Jeffery9707/VISIAR

Findings of the Association for Computational Linguistics: ACL 2025, pages 18384–18402

Developing an automatic visual story ideation system presents significant challenges. Such a system must demonstrate a deep understanding of visual assets, leverage open-world knowledge for creative story ideation, and efficiently navigate large collections of assets. The lack of established benchmarks or a well-defined framework further complicates this task. Current visual storytelling methods (Hong et al., 2023; Yang et al., 2023) often rely on predefined sequences of visual assets or accompanying text scripts, which are not available in real-world scenarios (such as creating short reels with a coherent story from hours of randomly taken vlogs), leaving the ideation process untouched.

To address these challenges, we propose **VISIAR**, **V**isual **I**deation through **S**equence **I**ntegration and **A**sset **R**earrangement, a framework that empowers MLLM for Visual Story Ideation. It encompasses three key stages: visual content understanding, candidate asset selection, and asset rearrangement for expressive storyline creation. In the first stage, MLLM is used to analyze visual content because of their remarkable visual understanding ability (Yin et al., 2024). However, MLLM's limitations of context window and long-context understanding make it unsuitable for the designated task. Therefore, in the second stage, we explore and implement multiple methods for candidate asset selection to empower MLLM to navigate through hundreds of visual assets. We develop feature-based clustering and graph-based clustering approaches. Specifically, we develop a group of **Story-Graph Enhanced** methods, which construct a story graph using MLLMs to propose candidate visual assets for each storyline. In particular, we invent the **LLM-Ideation Graph** method: Leveraging MLLM's open-world knowledge for story graph construction. Finally, in the third stage, MLLMs are employed to rearrange the selected assets and generate expressive storylines.

To establish a benchmark for this novel task, we curate a new dataset, **VTravel**, comprising short video clips to assemble realistic scenarios. User studies and the GPT-as-the-judge evaluation are conducted to assess our methods, demonstrating that our framework with the LLM-Ideation Graph method surpasses the performance of the state-of-the-art MLLM baseline in generating coherent and expressive visual stories.

In summary, our contributions are as follows:

- **Novel Task** We propose a novel and challenging task, Visual Story Ideation, which is underexplored yet of significant practical value for content creators.
- **Novel Framework** We introduce the first framework for visual story ideation, comprising three effective stages. In our framework, we propose innovative methodologies for constructing a visual story graph based on asset collections, empowering MLLMs to tackle the complex task of visual story ideation.
- **New Benchmark Dataset** To validate our framework and establish a robust benchmark, we have curated a novel dataset, called VTravel, which will be made publicly available upon publication to foster further research and innovation.

2 Related Work

To the best of our knowledge, we are the first to investigate the challenging Visual Story Ideation task. Our task is related to the visual storytelling task, ideation based on Large Language models, and tangentially related to Multi-modal Large Language models.

2.1 Visual Storytelling

Visual storytelling has been a popular research topic. The task usually involves generating a text story from a predefined sequence of images (Huang et al., 2016; Hong et al., 2023) or videos (Han et al., 2023). To improve story generation, Storyteller (ul Haque and Ghani, 2022) proposes an image caption method to generate story-type captions for images. (Yang et al., 2024) introduces generating video narrations with structured storylines with LLMs. These works play significant roles in visual storytelling. However, they all require a pre-defined sequence of visual assets, which provide guidance of the storyline and are the results from the ideation process. (Chen et al., 2019; Yang et al., 2023; Lu et al., 2023; Yturrizaga-Aguirre et al., 2022) provides another direction of visual storytelling by letting users illustrate a pre-defined story by retrieving clips from a large collection of images or videos through text script. (Sun et al., 2024) proposes a new dataset for Vision-Language Story Understanding, which supports the alignment between the story script and videos. (Gu et al., 2023) proposes TeVis, which retrieves an ordered sequence of images from a large dataset to visualize a high-level text synopsis. Script-to-Storyboard

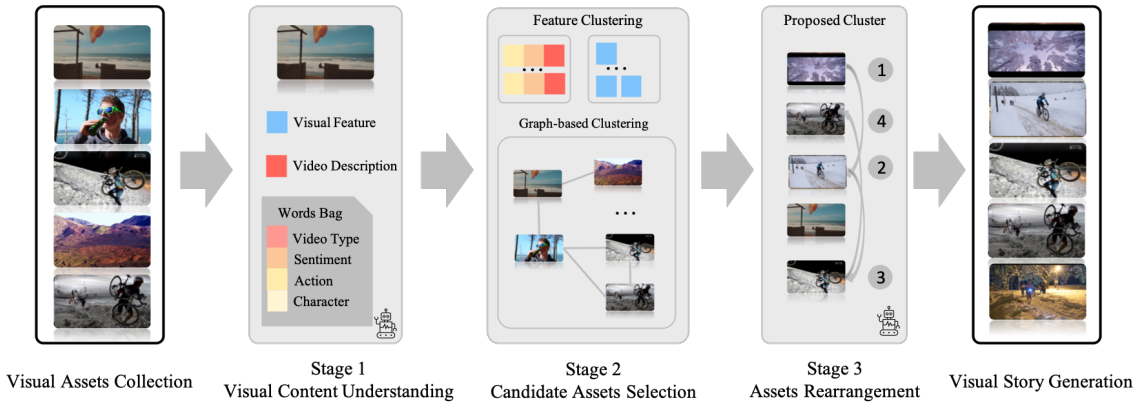


Figure 2: **Framework Overview**: Our framework VISIAR contains three stages. Visual Content Understanding: Utilize vision feature extractor and MLLM to extract features and text information from the visual assets. Candidate Assets Selection: Propose candidate assets for storylines through clustering methods. Assets Rearrangement: For each proposed cluster, use MLLMs to assemble the visual assets and generate a visual story.

pipeline (Tian et al., 2022; Rusu and Rusu, 2024) is also proposed to utilize text vision retrieval for storytelling. Those methods generally require a large collection of images or videos and an existing storyline. VSC (Choi et al., 2016) offers an alternative storytelling approach when the temporal information is available. It proposes to compose a visual story from a collection of clips with plot analysis. Further discussion and comparisons of temporal-based methods can be found in Appendix A.7. In summary, current research on visual storytelling does not provide insight into the story ideation process from existing videos or images. Our research fills in the gap between ideation and current visual storytelling works.

2.2 Ideation for Storytelling

Ideation is to assist content creators to develop expressive stories. It is challenging due to its exploratory and subjective nature. Multiple efforts have been put into this research direction. TaleStream (Chou et al., 2023) proposes to use tropes as an intermediate representation of stories to tackle this task. The system recommends tropes to help develop a text story. Metamorpheus (Wan et al., 2024) harnesses LLMs, such as ChatGPT, to help users relive their dream experience through visual story ideation. XCreation (Yan et al., 2023) also uses LLM to develop a Creativity Support Tool for storybook ideation. ScriptViz (Rao et al., 2024) provides a tool to assist scripting writing by retrieving frames from a movie dataset and revising the script. In summary, current ideation works usually provide a platform to assist users with storytelling, but no complete storyline is suggested automatically. Our visual story ideation focuses on video modality and provides a complete storyline without human interference.

2.3 Multi-modal Large Language Models

Multi-modal Large Language Models such as GPT-4o (Achiam et al., 2023) and LlaVa (Liu et al., 2024) bring new insight into Visual Story Ideation task because of their powerful capability. (Alayrac et al., 2022; Bai et al., 2023; Li et al., 2023; Driess et al., 2023; Wu et al., 2024; Wang et al., 2023; Zhu et al., 2023). However, the limitation of the context window size and the ability to understand long contexts impacts the visual story ideation performance. Therefore, we propose candidate asset selection in our framework to address this issue and empower the MLLM for the visual story ideation task.

3 Methodology

3.1 Method Overview

Given a collection of videos $\mathcal{V} = \{V_1, V_2, \ldots, V_n\}$, where each video V_i is represented as a sequence of frames $\{F_{i,t}\}$ indexed by time t, the goal is to generate a set of ordered "stories" $\mathcal{S} = \{S_1, S_2, \ldots, S_m\}$. Each story $S_j = \langle V_{j,1}, V_{j,2}, \ldots, V_{j,k} \rangle$ consists of an ordered subset of videos from $\mathcal{V}$ that together form a coherent and expressive visual storyline.

To leverage the ideation capabilities of Multi-modal Large Language Models (MLLMs) while addressing challenges such as limited context window size and the complexities of long-context understanding, our framework is designed with three stages: visual content understanding, candidate as-

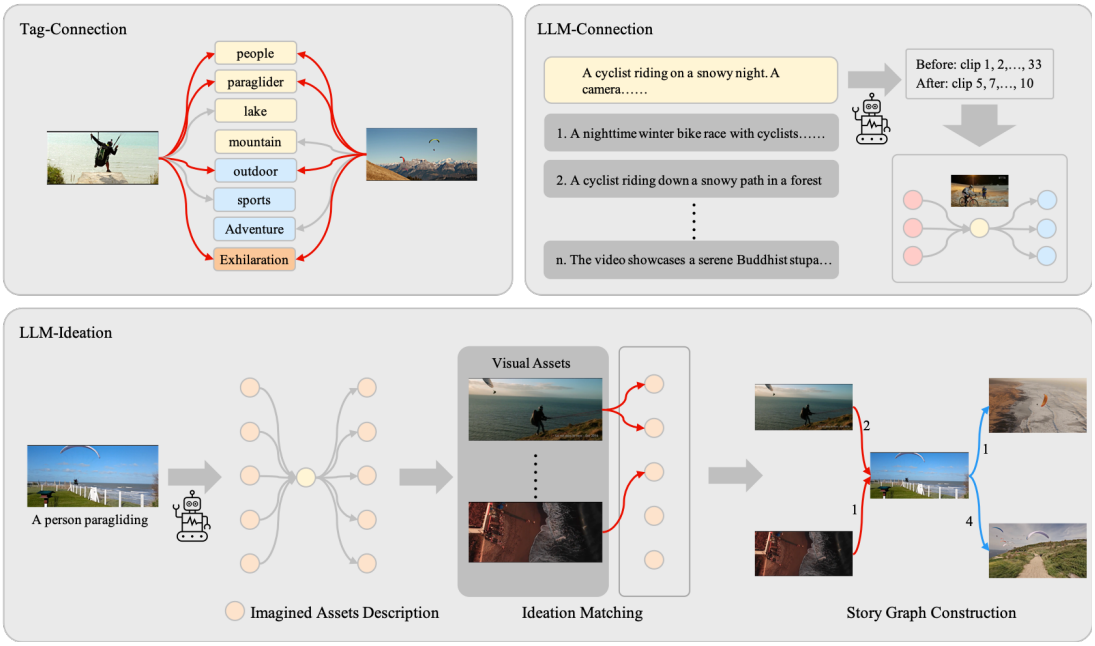


Figure 3: **Graph Connection Creation**: We experiment with three methods for creating the connection between assets. **Tag-Connection**: Define the edges to be the overlapped content word counts. **LLM-Connection**: Use LLM and all the descriptions of videos to create the edges. **LLM-Ideation**: Create the edges through ideation. Create imagined assets that connect to the input video. Then, the imagined assets are matched with existing videos by semantic similarity. Use the similarity to draw connections among visual assets so that the story graph is constructed.

sets selection, and assets rearrangement. Figure 2 shows an overview of our framwork.

In the first stage, keyframes are sampled from the input videos and transformed into textual representations. In the second stage, these features are utilized to design methods for selecting candidate visual assets. Finally, in the third stage, the selected assets are rearranged in a meaningful order, and an expressive storyline is automatically created with optional narratives.

3.2 Visual Content Understanding

We employ a multimodal large language model (MLLM), denoted as $\mathcal{M}$, to understand the visual content of videos. For each video V_i, we uniformly sample five frames, represented as $\{F_{i,t}\}_{t=1}^{5}$, where $F_{i,t}$ denotes the t-th sampled frame from video V_i. These frames are input into $\mathcal{M}$, which is prompted to perform the following tasks under the assumption that the frames are derived from a short video clip.

- *What is this video about?* A textual description of the video is extracted and denoted as D_i.
- *What kind of story can this video make?* Examples like *documentary*, *nature*, and *travel* guide the response, resulting in a video type list T_i.
- *What is the feeling from watching this video?* Sentiments are extracted under the assumption that similar emotions enhance coherence, represented as E_i
- *Who are the main characters in this video?* Main characters C_i are extracted, which may contribute to a cohesive storyline.
- *Extract action triplets* Action triplets, denoted as A_i, are extracted since they capture key content and aid in connecting clips.

Formally, we have:

$$\mathcal{M}(\{F_{i,t}\}_{t=1}^{5}, P_1) \rightarrow \{D_i, T_i, E_i, C_i, A_i\},$$

where P_1 represents the task-specific prompt. The output consists of D_i as sentences, T_i, E_i, C_i, and A_i as word bags.

3.3 Candidate Assets Selection

One approach to ideate a visual story from video collections is to input all sampled frames and descriptions into MLLMs and prompt them to perform the ideation task. While feasible for small

datasets, this becomes impractical for larger ones due to model limitations. Despite supporting up to $128k$ tokens, state-of-the-art MLLMs struggle with long contexts. Additionally, the computational cost associated with token usage poses another constraint.

To mitigate these issues and optimize MLLMs' capabilities, we introduce **candidate assets selection** as the second stage of our framework. This stage pre-selects visual assets that contribute to expressive, coherent stories, forming the basis for meaningful storylines. We explore multiple selection approaches, broadly categorized into feature-based clustering and graph-based methods.

3.3.1 Feature Clustering

The initiative of applying clustering methods over the collection of videos is that videos with similar content might be able to compose a visual story. We experimented with visual feature representation and semantic feature representation for this method.

Visual Clustering We uniformly sampled p frames from each video V_i. We utilize a pre-trained visual feature extractor, denoted as ϕ_v, to extract the visual feature. The visual representation W_v for V_i is computed as the mean of the extracted features from its sampled frames:

$$\mathbf{W^v}_i = \frac{1}{p}\sum_{t=1}^{p}\phi_v(F_{i,t}), \quad (1)$$

After extracting the visual representation, K-means (Hartigan and Wong, 1979) clustering is applied to propose the candidate visual assets for each storyline.

Text Clustering We utilize the extracted textual content to form a semantic representation for each video. T_i, E_i, C_i, and A_i are preprocessed to mitigate the effects of synonyms.

To generate textual representations, we employ a textual vectorizer, denoted as ϕ_t, which processes grouped single-word textual elements from T_i, E_i, C_i, and A_i to produce a frequency-based representation. For the video description D_i, we utilize a sentence embedding model, denoted as ϕ_s, to extract high-dimensional embeddings. The complete semantic representation for each video is then constructed as the concatenation of the textual vectorizer output and the embedding model output:

$$\mathbf{W^s}_i = [\phi_t(T_i, E_i, C_i, A_i), \phi_s(D_i)], \quad (2)$$

where $\mathbf{W^s}_i$ represents the semantic representation of video V_i. Finally, we apply K-means clustering on $\{\mathbf{W^s}_i\}$ to group videos based on their semantic content, generating clusters of candidate visual assets for each storyline.

We denote the obtained clusters of videos as $\mathcal{K} = \{K_1, K_2, \ldots, K_m\}$.

3.3.2 Graph-based clustering

A good visual story sequence does not necessarily consist of visually or semantically similar assets. Instead, it requires meaningful connections between the selected visual assets. To address this, we propose constructing a *story graph* that represents these connections using text features or insights from LLMs. Once the graph is constructed, graph clustering is applied to identify candidate assets for storytelling. Consider videos V_i as nodes in the story graph G, where the connection between V_i and V_j is represented as a weighted edge E_{ij}. To construct the graph and define these edges, we explore three approaches: **Tag-Based Graph**, **LLM-Connection Graph**, and **LLM-Ideation Graph**. Figure 3 provides an illustration of these methods.

Tag-Based Graph The weighted edge E_{ij} is determined by counting the number of overlapping words between T_i, E_i, C_i, A_i and T_j, E_j, C_j, A_j. To increase the diversity of events in the proposed candidate assets, we exclude action information from A_i and denote it as A_i^*. The weight on the edge is defined as:

$$E_{ij} = |(T_i, E_i, C_i, A_i^*) \cap (T_i, E_i, C_i, A_i^*)| \quad (3)$$

LLM-Connection Graph Multimodal large language models (MLLMs), denoted as $\mathcal{M}$, possess extensive knowledge and can effectively establish semantic connections between assets. To ensure the input remains within the context window size of $\mathcal{M}$, we provide only the video descriptions $\{D_i\}_{i=1}^n$, excluding visual content. For each video description D_i, $\mathcal{M}$ is prompted with the following query:

For the given video description D_i, which five videos can be placed before or after this video from the video collection $\{D_i\}_{i=1}^n$?

$$\mathcal{M}(\{D_i\}, P_2) \rightarrow \{V_k^i\}_{k=1}^{10},$$

where P_2 represents the task-specific prompt. The LLM provides connections by identifying up to 10 related videos $\{V_k^i\}_{k=1}^{10}$ for a given video V_i.

The weight E_{ij} is then defined as:

$$E_{ij} = \begin{cases} 1, & \text{if} \quad V_j \in \{V_k^i\}_{k=1}^{10}, \\ 0, & \text{otherwise.} \end{cases} \tag{4}$$

This process is repeated for each D_i, resulting in the construction of the story graph G.

LLM-Ideation Graph To leverage the exploratory and ideation capabilities of LLMs, we use **LLM-Ideation** to establish connections between V_i. For each video description D_i, we input it into the LLM and prompt the following task:

This is a clip taken from a short video. Imagine what possible q clips could happen before or after this clip. Give the description of those clips.

For each D_i, this process generates a collection of ideated video descriptions $\{I_r^i\}_{r=1}^{q}$, representing open-world knowledge. We denote P_3 as the task-specific prompt.

$$\mathcal{M}(\{D_i\}, P_3) \rightarrow \{I_r^i\}_{r=1}^{q},$$

To ground this knowledge in the existing video collection and establish connections, we calculate the cosine similarity between the embedding of ideated clips and the current clip collection. If the similarity is larger than a threshold θ, we increase the weight of the corresponding edge.

$$E_{ij} = \sum_{r=1}^{q} |\cos(\phi_s(I_r^i), \phi_s(D_j)) > \theta| \tag{5}$$

In this way, we build up the story graph G with open-world ideation knowledge.

Graph Clustering After constructing the story graph G, we apply graph clustering (Louvain) (Blondel et al., 2008), and obtain clusters of videos clusters of videos $\mathcal{K} = \{K_1, K_2, \ldots, K_m\}$ based on their connections with each other instead of similarity.

3.4 Assets Rearrangement

In the previous stage, we obtained clusters of videos $\mathcal{K} = \{K_1, K_2, \ldots, K_m\}$ using different methods. These clusters represent potential storylines. To produce the final visual story, we input the middle frame F_{mid} and the video description D_i of videos in K_m to the MLLM. We prompt the MLLM to select and arrange these videos and provide the video sequence for visual storytelling.

4 Experiments

4.1 Dataset

Ideally, we need a dataset containing hundreds of videos with diverse potential storylines. However, to the best of our knowledge, no datasets exist specifically for visual story ideation. To address this and to evaluate our methods, we curated a new dataset: VTravel.

VTravel We collected travel and nature videos from video platforms under CC-BY licenses. Using a scene detection model (PySceneDetect, 2024), we segmented videos into single-scene clips and manually filtered them, resulting in around 450 clips resembling a user's travel album. Each clip is 3–15 seconds long. Please refer to Appendix A.1 for more details about the construction of this dataset.

4.2 Models

We choose GPT-4o as the base MLLM $\mathcal{M}$ for this task due to its outstanding multimodal capabilities. With this, we build the baseline where we provide the model with the middle frame of all the videos and their description, prompting it to perform the visual ideation task. We include more details of the GPT-4o baseline in the Appendix A.3. For auxiliary models, especially for our proposed methods, we further use CLIP (Radford et al., 2021) as the visual feature extractor ϕ_v, TF-IDF vectorizer (Sparck Jones, 1972) as ϕ_t and Sentence-Bert (Reimers, 2019) as ϕ_s.

4.3 Qualitative Evaluation

To comprehensively evaluate each method, we first perform a qualitative analysis. Given the inherent challenges of directly comparing methods for the ideation task, we analyze each approach individually. Figures 4 illustrate the middle frame sequence of each clip. Some videos are truncated due to formatting limitations; additional results with full video visualized are provided in the appendix A.8.

GPT-4o Baseline When performing the visual story ideation task, GPT-4o turns to group visual assets with similar content together. It also appears to select irrelevant clips to form a short video, which leads to bad results (shown in the red rectangle).

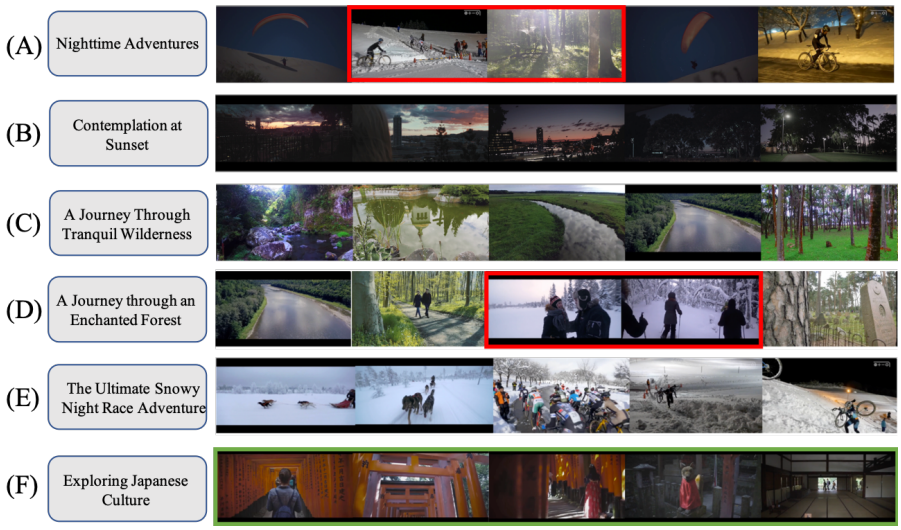


Figure 4: Qualitative results from our proposed **VTravel** dataset. The automatically generated topic is in the first column: The GPT-4o based baseline (A) displays disjoint content sets with an irrelevant clip in the middle (highlighted in red). Visual (B) and text (C) clustering methods produce monotonous and less engaging stories. The Tag-Connection graph (D) can create a progressive story but sometimes includes visually incoherent content (in red). The LLM-Connection graph (E) shares issues with methods B and C, resulting in less expressiveness. In contrast, the LLM-Ideation Graph (F) demonstrates superior quality, presenting a coherent yet engaging visual story.

Visual & Text Clustering The visual or text clustering methods propose visually or semantically similar assets as candidates for storytelling. We found that the generated videos are often composed of one single activity, resulting in a less expressive, monotonous storyline.

Tag-Connection Graph The Tag-Connection Graph method seems to be a happy middle between asset similarity and storyline development. For example, the resulting storyline (D) includes clips of rivers, forests, and tourism activities, illustrating a sense of progression in the storyline. However, the tag-based connectivity is relatively coarse, occasionally leading to the selection of clips that lack coherence within the visual story.

LLM-Connection Graph Utilizing LLM to suggest the connection between existing assets yields reasonable results. Although this method successfully formulates a storyline, it also tends to group similar assets for storytelling, which results in less expressive results. We suspect that the graph construction process may form cycles when assets focus on the same activity, contributing to these issues. More details are included in Appendix A.5.

LLM-Ideation Graph The LLM-Ideation Graph Method yields the best results, leveraging LLMs' open-world knowledge beyond existing video descriptions. The knowledge from LLM is grounded in the existing dataset and enriching storyline development. For instance, in Figure 4, the storyline (F) follows Japanese tourists exploring various landmarks, from statues to traditional Japanese buildings. Such coherent development of storylines is not observed in other methods. These findings suggest that our novel LLM-Ideation Graph Construction significantly enhances Visual Story Ideation and provides valuable insights into this challenging task. For more full video results, please refer to Appendix A.8.

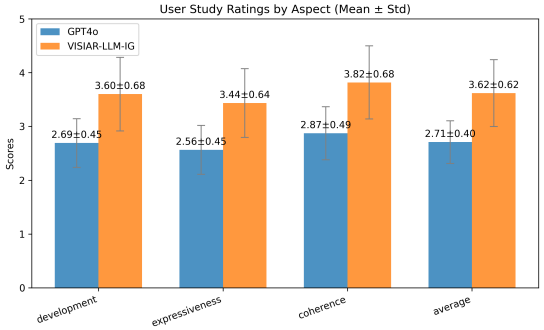


Figure 5: **User Study**: The VISIAR-LLM-IG method improves the performance in all aspects.

4.4 Quantitative Evaluation

4.4.1 User Study

Evaluating our proposed task quantitatively is challenging due to the subjective and complex nature

Metric	GPT-4o	Visual	Text	Tag-Graph	VISIAR-LLM-CG	VISIAR-LLM-IG
Development	3.2	3.5	3.5	3.5	3.6	**3.8**
Expressiveness	3.2	3.3	3.4	3.5	3.7	**3.8**
Coherence	3.3	3.6	3.2	3.4	**3.9**	**3.9**
Average	3.23	3.46	3.36	3.46	3.73	**3.83**

Table 1: GPT-4o Evaluation Scores Across Different Methods

of ideation. To validate the effectiveness of LLM-Ideation Graph, we compare it with GPT-4o baseline via user study. We denote this method as VISIAR-LLM-IG. Please note for this user study, pairwise comparison for different methods is not possible because the ideated sequence can be significantly different in terms of topic or selected clips among the methods. Instead, we focuses on the following attributes to assess the video results:

Development The storyline should demonstrate a clear progression, incorporating key elements from Freytag's Pyramid (Freytag, 1895), particularly the exposition (hook) and resolution (denouement). A rating of 5 indicates the presence of both a clear hook and resolution, while a rating of 1 indicates a lack of discernible structure.

Expressiveness While Freytag's Pyramid offers a useful structural framework for narratives, it does not inherently ensure that a story will be engaging or interesting, which is crucial for a visual story. Therefore, we ask users to rate it 1 if it appears dull and lacks transitions or exciting moments, and 5 if it overall captures meaningful and exciting scenes.

Coherence Another requirement for a visual story is that all clips should visually complement each other and share a consistent theme. Therefore, we ask users to rate coherence from 1 to 5, with each lower rating indicating a higher proportion of clips that are out of place (for example, a rating of 2 might indicate 3 or more scenes are out of place).

With these metrics, we conducted a user study to compare the performance of the GPT-4o baseline with our LLM-Ideation Graph-based method. The participants are 15 volunteers who are familiar with the video media. The results of the user study are presented in Figure 5. As shown, our method significantly improves ideation performance across all aspects, surpassing GPT-4o by 33.5%. We conduct Paired t-test and Wilcoxon signed-rank test, showing that the improvements are statistically significant ($p < 0.01$). More details of the user study, including an agreement analysis, are included in Appendix A.4

4.4.2 GPT-as-the-judge

To supplement human evaluation, we implemented a *GPT-as-the-Judge* evaluation, which is widely adopted in works like (Li et al., 2024; Xiong et al., 2024; Dubois et al., 2024; Hu et al., 2023). GPT-4o was prompted to rate the videos based on our survey criteria. The results are in table 1. The prompt is attached in Appendix A.6

These results show a strong alignment between GPT-4o's evaluation and the user study, particularly in comparing GPT-4o baseline vs. LLM-Ideation. The overall trends further confirm our observations in Section 4.3. To evaluate the effectiveness of GPT as a judge, we include an agreement analysis with the user study and present additional experiments with statistical testing in Appendix A.5.

The visual clustering approach exhibits high coherence due to its focus on visual consistency, whereas textual clustering, lacking this focus, results in lower coherence. Both methods fail to produce expressive stories with satisfying development. The tag-graph approach attempts to balance expressiveness and coherence but remains suboptimal, as tag information is often low-level and coarse.

By incorporating the extensive open-world knowledge and reasoning capabilities of large language models (LLMs), both LLM-Connection and LLM-Ideation approaches improve all evaluation metrics. Notably, the LLM-Ideation approach achieves the best performance, yielding the highest scores in development and expressiveness while maintaining a high coherence score. These findings highlight the effectiveness of our LLM-Ideation method.

5 Conclusion

In this paper, we introduce **Visual Story Ideation**, a novel task focused on selecting and combining

visual asset sequences to create compelling storylines from limited collections. Unlike prior work relying on predefined sequences or scripts, our approach automates the ideation process, addressing a key gap in visual storytelling.

For this task, we also propose **VISIAR**, **V**isual **I**deation through **S**equence **I**ntegration and **A**sset **R**earrangement, a framework leveraging state-of-the-art Multimodal Large Language Models (MLLMs), and a Story-Graph-Enhanced Selection process. Evaluations on a new dataset, VTravel, demonstrate its ability to outperform baselines by generating coherent and expressive asset sequences.

Our work establishes a foundation for future research in automated visual story ideation, with potential applications in advertising, media production, and creative content generation. We believe this framework opens new avenues for innovation in both academic research and practical implementations of visual storytelling.

6 Limitation

Our framework is promising and provides new insights into MLLM research. There is still space for improvement in terms of story graph construction. For example, when utilizing the LLM-Ideation graph construction method, some assets might not gain enough connectivity with other assets. The story graph might not have a nice structure for clustering in rare cases. We aim to address this limitation in future research.

We also want to address the fact that our method aims to ideate storylines from the asset collection through MLLM automatically. It is possible that the ideated story inherited the biases from the MLLM, which could be problematic. Future investigation on unbiased MLLMs will likely address this issue.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. 2022. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*.

Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. 2008. Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, 2008(10):P10008.

Shizhe Chen, Bei Liu, Jianlong Fu, Ruihua Song, Qin Jin, Pingping Lin, Xiaoyu Qi, Chunting Wang, and Jin Zhou. 2019. Neural storyboard artist: Visualizing stories with coherent image sequences. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 2236–2244.

Jinsoo Choi, Tae-Hyun Oh, and In So Kweon. 2016. Video-story composition via plot analysis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3122–3130.

Jean-Peïc Chou, Alexa Fay Siu, Nedim Lipka, Ryan Rossi, Franck Dernoncourt, and Maneesh Agrawala. 2023. Talestream: Supporting story ideation with trope knowledge. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, pages 1–12.

Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, et al. 2023. Palm-e: An embodied multimodal language model. *arXiv preprint arXiv:2303.03378*.

Yann Dubois, Balázs Galambosi, Percy Liang, and Tatsunori B Hashimoto. 2024. Length-controlled alpacaeval: A simple way to debias automatic evaluators. *arXiv preprint arXiv:2404.04475*.

Gustav Freytag. 1895. *Technique of the drama: An exposition of dramatic composition and art*. S. Griggs.

Xu Gu, Yuchong Sun, Feiyue Ni, Shizhe Chen, Xihua Wang, Ruihua Song, Boyuan Li, and Xiang Cao. 2023. Tevis: Translating text synopses to video storyboards. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 4968–4979.

Mingfei Han, Linjie Yang, Xiaojun Chang, and Heng Wang. 2023. Shot2story20k: A new benchmark for comprehensive understanding of multi-shot videos. *arXiv preprint arXiv:2312.10300*.

John A Hartigan and Manchek A Wong. 1979. Algorithm as 136: A k-means clustering algorithm. *Journal of the royal statistical society. series c (applied statistics)*, 28(1):100–108.

Xudong Hong, Asad Sayeed, Khushboo Mehra, Vera Demberg, and Bernt Schiele. 2023. Visual writing prompts: Character-grounded story generation with curated image sequences. *Transactions of the Association for Computational Linguistics*, 11:565–581.

Yushi Hu, Benlin Liu, Jungo Kasai, Yizhong Wang, Mari Ostendorf, Ranjay Krishna, and Noah A Smith. 2023. Tifa: Accurate and interpretable text-to-image faithfulness evaluation with question answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20406–20417.

Ting-Hao K. Huang, Francis Ferraro, Nasrin Mostafazadeh, Ishan Misra, Jacob Devlin, Aishwarya Agrawal, Ross Girshick, Xiaodong He, Pushmeet Kohli, Dhruv Batra, et al. 2016. Visual storytelling. In *15th Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL 2016)*.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR.

Tianle Li, Wei-Lin Chiang, Evan Frick, Lisa Dunlap, Tianhao Wu, Banghua Zhu, Joseph E Gonzalez, and Ion Stoica. 2024. From crowdsourced data to high-quality benchmarks: Arena-hard and benchbuilder pipeline. *arXiv preprint arXiv:2406.11939*.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2024. Visual instruction tuning. *Advances in neural information processing systems*, 36.

Yu Lu, Feiyue Ni, Haofan Wang, Xiaofeng Guo, Linchao Zhu, Zongxin Yang, Ruihua Song, Lele Cheng, and Yi Yang. 2023. Show me a video: A large-scale narrated video dataset for coherent story illustration. *IEEE Transactions on Multimedia*.

PySceneDetect. 2024. PySceneDetect. Accessed: 2024-12-14.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Anyi Rao, Jean-Peïc Chou, and Maneesh Agrawala. 2024. Scriptviz: A visualization tool to aid scriptwriting based on a large movie database. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology*, pages 1–13.

N Reimers. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Adrian Rusu and Amalia Rusu. 2024. Script-to-storyboard-to-story reel framework. In *2024 28th International Conference Information Visualisation (IV)*, pages 350–355. IEEE.

Karen Sparck Jones. 1972. A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*, 28(1):11–21.

Yidan Sun, Jianfei Yu, and Boyang Li. 2024. Multilingual synopses of movie narratives: A dataset for vision-language story understanding. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 13488–13504.

Xi Tian, Yongliang Yang, and Qi Wu. 2022. Script-to-storyboard: A new contextual retrieval dataset and benchmark. *Computational Visual Media*, 3(4):5.

Anwar ul Haque and Sayeed Ghani. 2022. The storyteller: Computer vision driven context and content generation system.

Qian Wan, Xin Feng, Yining Bei, Zhiqi Gao, and Zhicong Lu. 2024. Metamorpheus: Interactive, affective, and creative dream narration through metaphorical visual storytelling. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 1–16.

Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, et al. 2023. Cogvlm: Visual expert for pretrained language models. *arXiv preprint arXiv:2311.03079*.

Junda Wu, Zhehao Zhang, Yu Xia, Xintong Li, Zhaoyang Xia, Aaron Chang, Tong Yu, Sungchul Kim, Ryan A Rossi, Ruiyi Zhang, et al. 2024. Visual prompting in multimodal large language models: A survey. *arXiv preprint arXiv:2409.15310*.

Tianyi Xiong, Xiyao Wang, Dong Guo, Qinghao Ye, Haoqi Fan, Quanquan Gu, Heng Huang, and Chunyuan Li. 2024. Llava-critic: Learning to evaluate multimodal models. *arXiv preprint arXiv:2410.02712*.

Zihan Yan, Chunxu Yang, Qihao Liang, and Xiang'Anthony' Chen. 2023. Xcreation: A graph-based crossmodal generative creativity support tool. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, pages 1–15.

Dingyi Yang, Chunru Zhan, Ziheng Wang, Biao Wang, Tiezheng Ge, Bo Zheng, and Qin Jin. 2024. Synchronized video storytelling: Generating video narrations with structured storyline. *arXiv preprint arXiv:2405.14040*.

Guoxing Yang, Haoyu Lu, Zelong Sun, and Zhiwu Lu. 2023. Shot retrieval and assembly with text script for video montage generation. In *Proceedings of the 2023 ACM International Conference on Multimedia Retrieval*, pages 298–306.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. 2024. A survey on multimodal large language models. *National Science Review*, page nwae403.

Arian Yturrizaga-Aguirre, Camilo Silva-Olivares, and Willy Ugarte. 2022. Story visualization using image-text matching architecture for digital storytelling. In *2022 IEEE Engineering International Research Conference (EIRCON)*, pages 1–4. IEEE.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*.

A Appendix

A.1 Dataset

A.1.1 Dataset Construction

VTravel We collected travel and nature videos from video platforms under CC-BY licenses. We use keywords such as 'Travel' and 'Nature' to search the videos. We include videos of various events and stories to assemble a personal album for traveling.

5 Human annotators select long videos with multiple scenes about diverse travel-related events. Most long videos are around 5 minutes. After applying scene detection (PySceneDetect, 2024) and video segmentation on those videos, we obtained around 600 short clips. The annotators filter videos based on the following:

- **Quality control**: Removal of low-quality videos, including blurry clips due to fast movements, meaningless content, or those with distracting visual effects.
- **Content safety** Exclusion of inappropriate material, such as nudity or content suggesting violence.
- **Relevance** Ensuring alignment with travel storytelling.

As a result, around 450 video clips are selected to construct the VTravel dataset, resembling a user's travel album. This process ensures that VTravel serves as a diverse and feasible dataset for visual story ideation. Each clip is 3–15 seconds long. We resize them to 1280×720 resolution. Figure 6 shows a snippet of the dataset. To further analyze the diversity of the dataset, we apply LDA topic modeling and present the identified topics using representative keywords in Table 2. As illustrated in the chart, our dataset spans various themes including nature, events, and human activities, offering broad visual coverage.

A.1.2 Dataset Comparison

To the best of our knowledge, we are the first to explore the visual story ideation task and collect the corresponding dataset. The construction process of the VIST dataset shares some similarities with our task. However, we would like to clarify the fundamental differences between the problem setting of VIST and ours:

Topic	Keywords	% Clips
0	street, lights, scene	5.63%
1	person, trees, blue	4.05%
2	flying, birds, red	1.13%
3	view, aerial, landscape	7.43%
4	snowy, night, snow	10.59%
5	person, paragliding, ocean	15.09%
6	person, sky, paragliding	6.31%
7	scene, buildings, swimming	2.93%
8	forest, trees, serene	6.53%
9	wooden, person, traditional	2.70%
10	forest, showing, aerial	2.70%
11	shots, video, person	2.70%
12	forested, showcasing, spectators	1.35%
13	bustling, person, cityscape	2.48%
14	video, bicycle, person	4.28%
15	green, lush, landscape	3.83%
16	area, scene, beach	4.73%
17	view, buildings, aerial	7.43%
18	colorful, narrow, close	4.73%
19	dunes, night, sand	3.38%

Table 2: LDA Topic Modeling (Top 20 Topics)

VIST Dataset Characteristics

- Uses pre-filtered Flickr albums (10–50 images per album).
- Each album has a predefined topic (e.g., parties, offices).
- Albums are curated with a 48-hour time constraint and a coherent storyline.

Therefore, an album is already a collection of good candidates for a storyline with possibly predefined topics.

Our Project Characteristics

- We process large-scale, diverse visual assets (hundreds of images/videos).
- No predefined topics—multiple stories must be ideated from broad visual data.
- The task is significantly more challenging due to increased variability.

Using VIST for our task faces the following challenges:

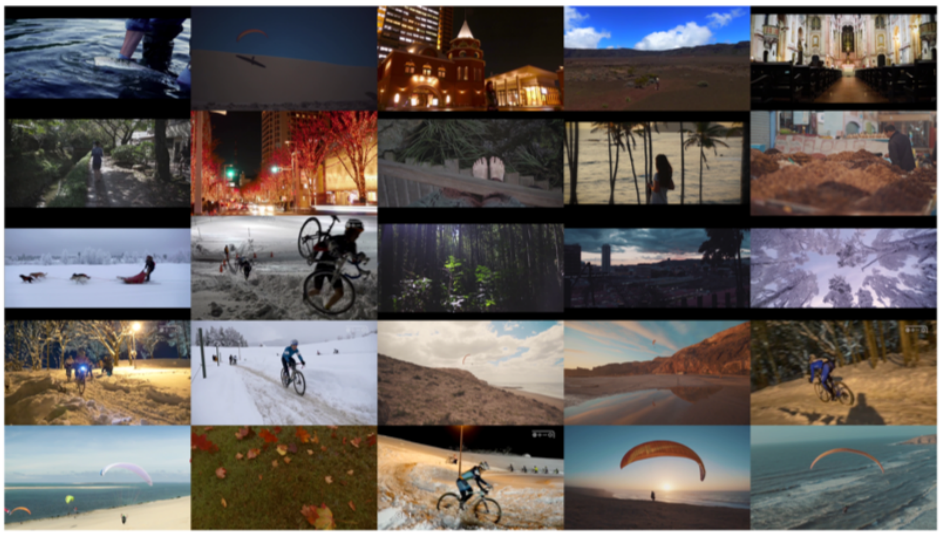


Figure 6: **VTravel** Here is a snippet of the VTravel Dataset, which contains a variety of events and scenes about nature and travel.

Limited Story Ideation Scope : Within each album, the topic of the story is pre-defined because all pictures are from the same category. The number of pictures is also limited to 50. This constrains creativity and limits the open-ended ideation our project aims for.

Lack of Cross-Album Annotations : Combining images from multiple albums could result in meaningful stories, but VIST does not annotate such possibilities. This makes it unsuitable as a direct resource for ideation.

Thus, while VIST is valuable for storytelling research, it is not directly applicable to our problem setting.

We also would like to point out that it is challenging to annotate the 'ground truth' of the visual story ideation task, as there could be countless combinations.

A.2 Experiment Setting

For the feature-based clustering method, we apply the K-means clustering method. The parameter k, which determines the number of clustering, is set to be the number of videos divided by 15. We assume for each cluster, there should be around 15 videos.

For Graph Clustering, we empirically set the resolution to be 1.8. For the LLM-Ideation Graph method, the threshold for determining whether to add an edge is set to 0.5.

We report the approximate token usage of GPT-4o for the visual understanding stage. For all experiments, we use low-resolution images as input. Each video costs around 1500 tokens for text extraction.

A.3 GPT4o Baseline

We provide additional details of the GPT4o baseline here. The model receives the middle frame and text description for each video. And the following prompt is used:

> Here are descriptions of several videos along with one frame from each. Make multiple stories by selecting and combining some of the videos based on the text description and visual features.
>
> For each story, it should contain no more than 10 videos.
>
> You only need to output the top 20 video sequences that make the best stories.
>
> Output the theme of the story in this group. Also output the orders of the selected videos for each story.
>
> Answer with a JSON object with group id from 0 to the total number of groups as the key and the answer as the value.
>
> Here is an example of the output JSON con-

tent:
"0": "theme": "City at Night", "orders": ["1", "4", "9", "15", "16", "24", "38", "51", "99", "132"]

A.4 User Study

A.4.1 Survey Setup

Evaluating ideation methods is very challenging. Due to the exploratory nature of the ideation, pair comparison is also impossible because the ideated sequence can be significantly different in terms of topic or selected clips. Therefore, we conducted a user study based on volunteer user participation and asked the users to give ratings on 'Development,' 'Expressiveness,' and 'Coherence.' The users are mostly college students who are familiar with the Vlog or other types of short videos. We utilize the following setup to strengthen the credibility of the evaluation.

- **User Study Design**: We recruited volunteers who are familiar with visual storytelling and hold a Master's degree or higher. Each participant received detailed instructions to align with the survey's purpose. We also provide short training sessions to help participants understand the metrics and survey details.
- **Evaluation Metrics**: We defined three key metrics—development, expressiveness, and coherence—for annotators to assess.
- **Multiple Annotations**: Each sample was annotated by at least 15 annotators to ensure reliability. The final score was calculated as the average of their annotations.

Part of the user study survey and the instructions are illustrated below:

This is a survey for the visual story ideation project. You will be asked to rate 20 videos. Each video is less than one minute. You will be asked to rate the videos regarding (1) Story development. (2) Expressiveness (3) Coherence.

Instructions:

(1) Story development (The title of the video suggests the topic of the video. Do you agree that there is such a storyline?)

Rating from 1 to 5.

1: I cannot identify the storyline.

3: I agree that this video is trying to tell a story. But the semantic development is bad.

5. There is a storyline, and I can identify most of the development moments in the story.

(2) Expressiveness (How expressive do you think the story of this video is?)

Rate from 1 to 5:

1. The story is nonsense

3. The story exists but boring. The visual diversity is low.

5. The story exists and is expressive

(3) Coherence (Do you feel all the clips belong to this video?)

Rate from 1 to 5:

1. The clip combination is meaningless

3. One or two clips seem to be strange. But I can still tell the story from the video.

5. All clips are coherent in terms of storytelling.

A.4.2 User Study Results

We provide an agreement analysis and show the results in the table 3. These values fall within the fair-to-moderate $(0.35, 0.55)$ range, which is reasonable given the subjective nature of the task and variability among human raters.

In order to further evaluate the user study results, we performed the paired t-tests and Wilcoxon signed-rank tests for the within-subject statistical tests. See table 4 for more details. The results consistently indicate significant differences ($p < 0.01$) for all individual aspects as well as the combined average. This consistency across tests reinforces the reliability of the findings.

Aspect	ICC Score
Development	0.402
Expressive	0.350
Coherence	0.553
Average	0.506

Table 3: User Agreement scores by aspect

A.5 Additional Analysis

We provide additional analysis of the impact of the story graph on candidate selection, particularly of the LLM-Connection method.

We observed the LLM-Connection graph often links similar assets, leading to dense intra-cluster connections.

Consider the following example:

- (A) A nighttime city scene adorned with festive lights.

Aspect	t-value	p-value (t-test)	W-value	p-value (Wilcoxon)
Development	4.964	0.0006	2.000	0.0029
Expressiveness	5.528	0.0003	0.000	0.0050
Coherence	4.440	0.0013	1.500	0.0029
Average	5.157	0.0004	1.000	0.0020

Table 4: Paired t-test and Wilcoxon signed-rank test results for each aspect and the combined average between GPT4o and VISIAR-LLM-IG.

- (B) A nighttime scene featuring beautifully illuminated trees.
- (C) A vibrant display of illuminated trees and festive decorations.

The LLM connected all those assets to each other. The clustering process may result in less expressive storylines due to redundant asset selection.

A.6 GPT-as-the-judge

A.6.1 Agreement Analysis

To verify the effectiveness of GPT-as-the judge and the alignment between the automatic evaluation and the human evaluation, we provide an agreement analysis on GPT4o vs Human. The results are in table 5.

This demonstrates strong alignment between GPT and human assessments, supporting the use of GPT as an evaluator.

Aspect	ICC (GPT vs Human)
Development	0.685
Expressive	0.522
Coherence	0.811
Average	0.768

Table 5: GPT and Human Agreement ratings by aspect

A.6.2 Significance Testing

We ran GPT evaluations 20 times per clip. Table 6 reports the mean value of the testing:

With all the ratings obtained, we performed the Paired t-test and Wilcoxon signed-rank test for the `llm_ideation` method against all other methods. The results are shown in Table 7:

The improvements shown by `llm_ideation` are statistically significant over all baselines.

A.6.3 Prompt

We include the prompt used for GPT4o to evaluate the generated video sequence here:

This is a survey for the visual story ideation project. You will be asked to rate a video. The video is combined from some short video clips. And you will be given one frame form each clip. The order of the clips matters for the visual story.

You will be asked to rate the videos regarding (1) Story development. (2) Expressiveness (3) Coherence.

Instructions:

(1) Story development

[The title of the video suggests the topic of the video. Do you agree that there is such a storyline?]

Rating from 1 to 5.

1: I cannot identify the storyline.

3: I agree that this video is trying to tell a story. But the semantic development is bad.

5. There is a storyline and I can identify most of the development moments of the story.

(2) Expressiveness

[How expressive do you think the story of this video is?]

Rate from 1 to 5:

1. The story is non-sense.

3. The story exists but boring. The visual diversity is low.

5. The story exists and is expressive.

(3) Coherence

[Do you feel all the clips belong to this video?]

1. The clip combination is meaningless

3. One or two clips seem to be strange. But I can still tell the story from the video.

5. All clips are coherent in terms of storytelling.

Answer with a JSON object with metric names as the key and the rating as the value. Here is an example of the output json file's content.

Method	Development	Expressive	Coherence	Overall
gpt4o_baseline	3.1600	3.3400	3.1950	3.2317
llm_connection	3.6600	3.7150	3.8350	3.7367
llm_ideation	**3.8250**	**3.8800**	**3.9750**	**3.8933**
tag_graph	3.4650	3.4750	3.5150	3.4850
text_sim	3.4300	3.4400	3.1550	3.3417
visual_sim	3.4650	3.4550	3.5450	3.4883

Table 6: Mean GPT evaluation scores per method

Comparison	Aspect	t-test (t, p)	Wilcoxon (w, p)
gpt4o_baseline vs llm_ideation	Development	t = 9.6615, p < 0.0001	w = 1787.0000, p < 0.0001
	Expressive	t = 8.3195, p < 0.0001	w = 1517.0000, p < 0.0001
	Coherence	t = 7.5365, p < 0.0001	w = 3401.5000, p < 0.0001
	Overall	t = 9.0306, p < 0.0001	w = 2646.5000, p < 0.0001
llm_connection vs llm_ideation	Development	t = 4.3389, p < 0.0001	w = 312.0000, p < 0.0001
	Expressive	t = 4.8771, p < 0.0001	w = 154.0000, p < 0.0001
	Coherence	t = 2.0048, p = 0.0463	w = 1883.0000, p = 0.0591
	Overall	t = 3.9211, p < 0.0001	w = 1825.0000, p < 0.0001
tag_graph vs llm_ideation	Development	t = 6.2024, p < 0.0001	w = 2057.0000, p < 0.0001
	Expressive	t = 9.9308, p < 0.0001	w = 322.0000, p < 0.0001
	Coherence	t = 4.8122, p < 0.0001	w = 3323.5000, p < 0.0001
	Overall	t = 6.9154, p < 0.0001	w = 3197.5000, p < 0.0001
text_sim vs llm_ideation	Development	t = 7.2606, p < 0.0001	w = 2077.0000, p < 0.0001
	Expressive	t = 9.9853, p < 0.0001	w = 654.0000, p < 0.0001
	Coherence	t = 11.1432, p < 0.0001	w = 562.0000, p < 0.0001
	Overall	t = 10.8082, p < 0.0001	w = 2070.0000, p < 0.0001
visual_sim vs llm_ideation	Development	t = 6.2024, p < 0.0001	w = 1358.0000, p < 0.0001
	Expressive	t = 8.6144, p < 0.0001	w = 658.0000, p < 0.0001
	Coherence	t = 4.9444, p < 0.0001	w = 3326.0000, p < 0.0001
	Overall	t = 6.9863, p < 0.0001	w = 3226.0000, p < 0.0001

Table 7: Paired t-test and Wilcoxon signed-rank test results comparing `llm_ideation` with other methods. Statistically significant results are indicated by $p < 0.0001$ unless otherwise noted.

{ "Development": 4, "Expressiveness": 3, "Coherence": 2 }

Please proceed with the story sequence. The sequence is provided by video frames and descriptions in the correct order.

A.7 Temporal Baseline Discussion

Our method does not rely on temporal information, making it applicable to a wider range of use cases. To further demonstrate its effectiveness, we compare it with a naive temporal baseline.

We simulate a sequence of video clips captured by a user during a single activity, each annotated with a timestamp. The naive temporal baseline stitches the clips together in chronological order, following the approach suggested in (Choi et al., 2016).

We apply both our method and the temporal baseline, with the results shown in Figure 7.

As illustrated, the temporally ordered sequence begins with clips of human activities, followed by scenes of the surroundings, as the users turn to focus on the human first and then appreciate the environment. This simple chronological stitching fails to produce a coherent or expressive story. In contrast, our approach yields more compelling and satisfying results without assuming any temporal structure.

A.8 Additional Results

We included more results in this section. For each ideated story, MLLM will automatically generate the theme of the story and use it as the title. See Figure 8, 9, 10, 11, 12, 13 for the additional results.

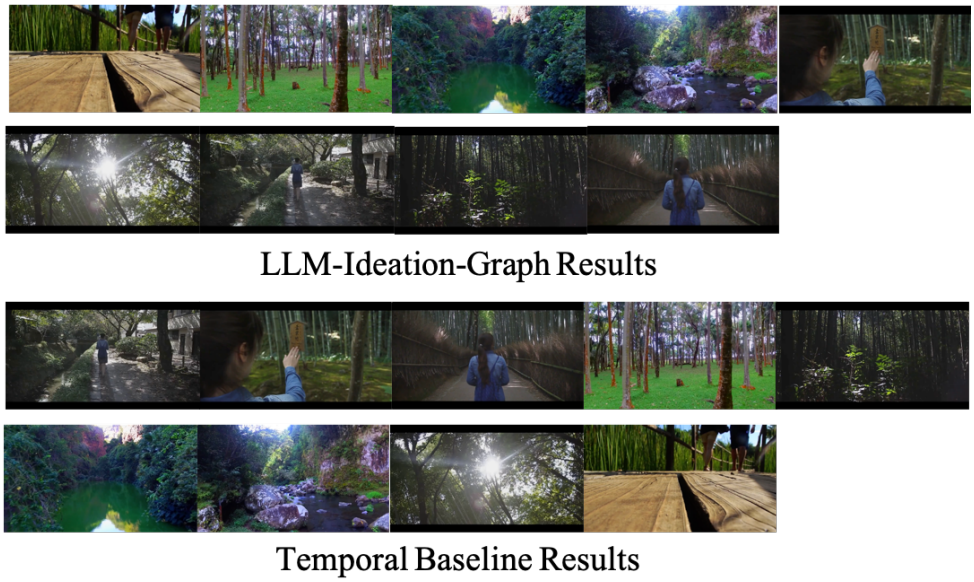


Figure 7: Results for LLM-Ideation Graphs and simulated temporal baseline

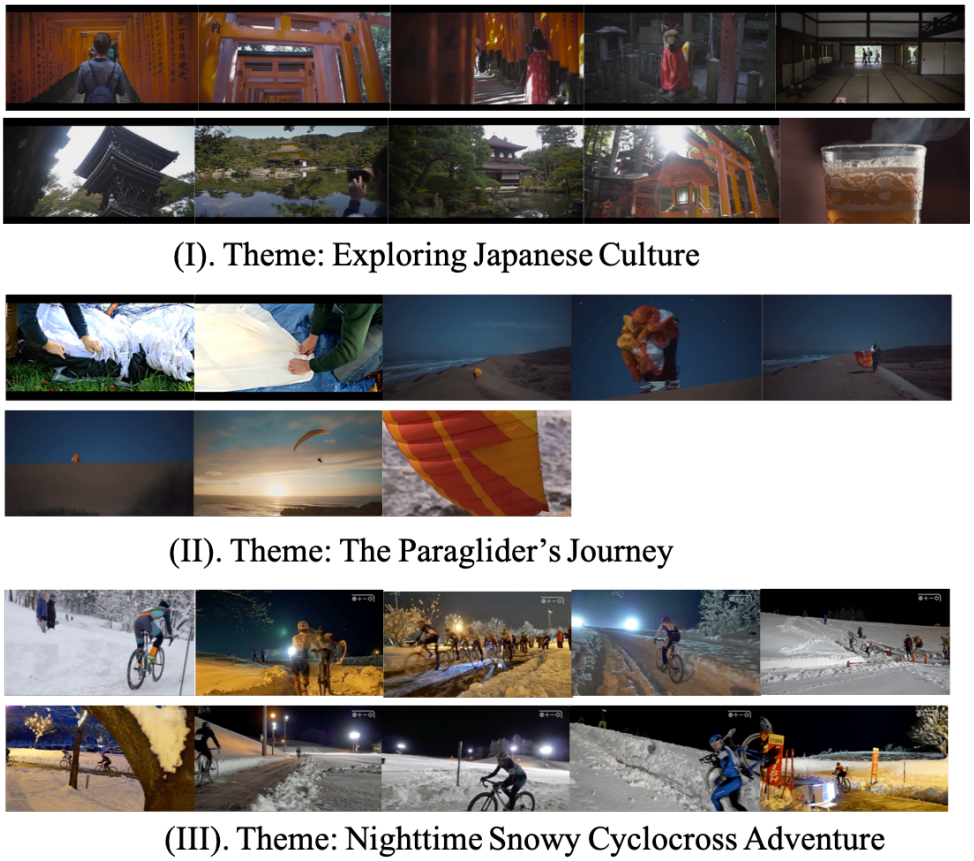
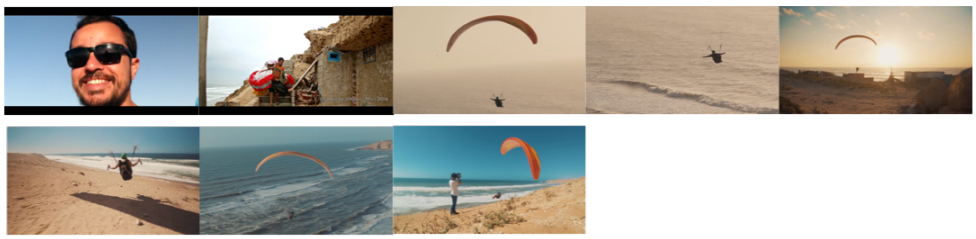
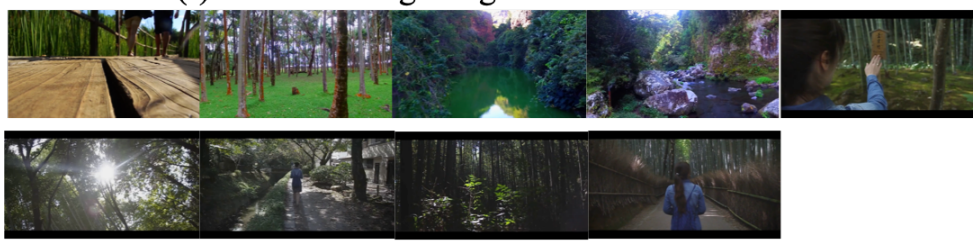


Figure 8: **Results for LLM-Ideation Graph**: Full results including all the clips

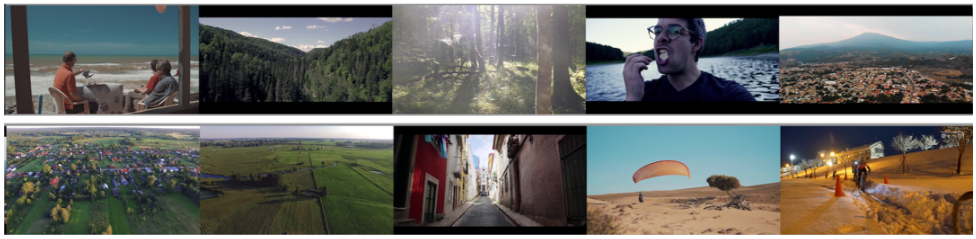


(I). Theme: Paragliding Adventure



(II). Theme: Journey Through Nature

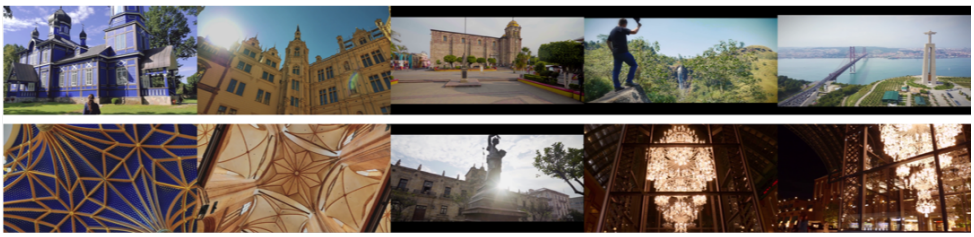
Figure 9: **Results for LLM-Ideation Graph**: Full results including all the clips



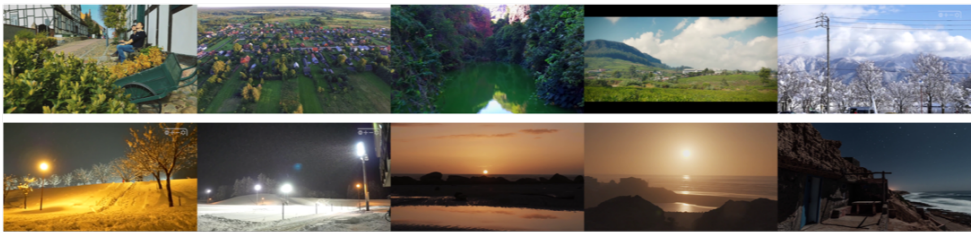
(I). Theme: Adventures in Nature

(II). Theme: Serene Landscapes

Figure 10: **Results for GPT4o**: Full results including all the clips



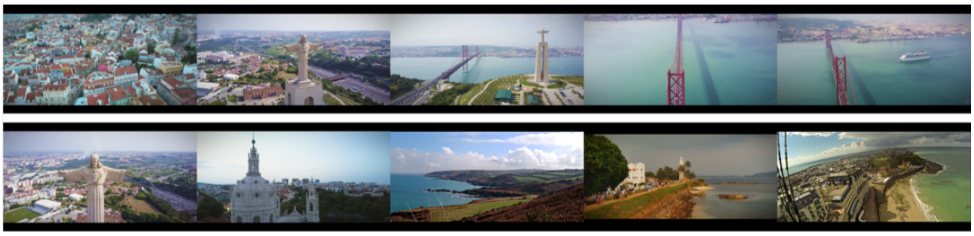
(I). Theme: Journey through Architecture and Nature



(II). Theme: Journeys Through Serenity and Seasons

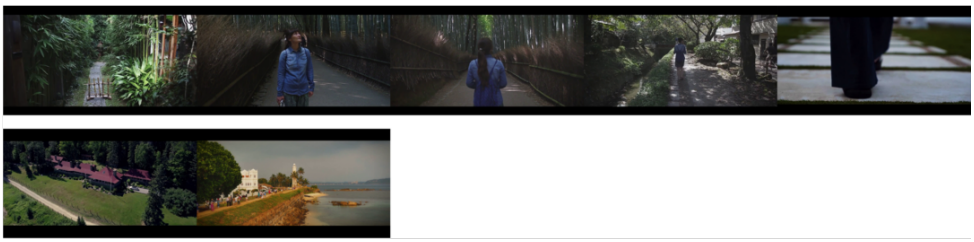
Figure 11: **Results for Text Similarity Clustering**: Full results including all the clips

(I). Theme: Exploring the Vibrant Streets



(II). Theme: Coastal City Exploration

Figure 12: **Results for Visual Similarity Clustering**: Full results including all the clips

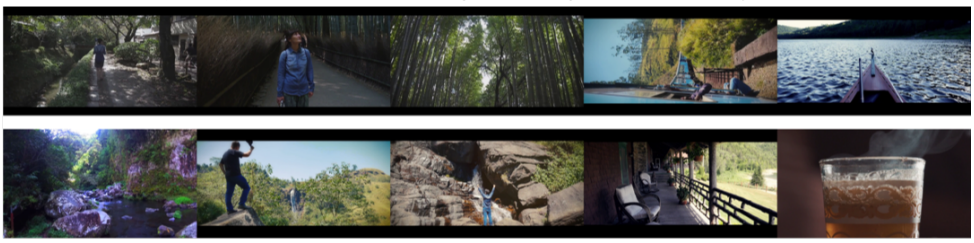


(I). Theme: A Journey Through Different Paths

(II). Theme: A Paraglider's Journey Across Diverse Landscapes

Figure 13: **Results for Tag Graph Method**: Full results including all the clips

(I). Theme: A Magical Night in the City



(II). Theme: A Journey Through Nature's Serenity

Figure 14: **Results for LLM-Connection Graph**: Full results including all the clips