DreamFrame: Enhancing Video Understanding via Automatically Generated QA and Style-Consistent Keyframes

This supplementary document is organized as follows:

- Section A contains more details about the model architecture of DreamFrame-7B.
- Section B contains more details about training parameters.
- Section C contains prompt examples for the stage one of our approach.
- Section D contains type distribution and more examples of our generated data.
- Section E contains comparison results on movie understanding.

A MORE DETAILS ON MODEL ARCHITECTURE

Here, we introduce the fundamental modules used in LLaMA-VID. LLaMA-VID basically consists of a visual encoder, a text decoder, a projector and a LLM. For visual encoder, LLaMA-VID use EVA-G [2] as its ViT-based backbone and the patch size is set to 14. For text decoder, Qformer-7b [1] is used in our experiment. For projector, one-layer MLP is used to transform the embedding into context token. For LLM, we use vicuna-7b.

B MORE TRAINING DETAILS ON DREAMFRAME-7B

As mentioned, we conduct experiments after the second stage of LLaMA-VID. During the model training phase, we employe the original LLaMA-VID configuration as the foundation for our training process. We utilize 2 NVIDIA A100 GPUs. To conserve GPU memory, we employ deepspeed with zero3 during model training, disabling tf32 and opting for fp16. The remaining parameters are shown in Table A

Settings	Parameter
Batch size	4
Learning rate	2e-5
Learning schedule	Cosine decay
Warmup ratio	0.03
Weight decay	0
Optimizer	AdamW
Vision encoder	Freeze
Text decoder	Freeze
Max token	65536

Table A: Training settings.

C PROMPT EXAMPLE

We illustrate the movie plot generation prompt in Figure A. By providing specific elements such as themes, overview, and styles, we guide GPT-4 to produce movie-level key frame descriptions tailored to the latter generation process.

D MORE DETAILS AND RESULTS OF GENERATED DATA

We show more results of our generated data in Figure B and Figure C.

E COMPARISON RESULTS ON MOVIE UNDERSTANDING

We demonstrate several comparison results for the comprehension of movie clips to show how we improve our baseline model LLaMA-VID in Figure D. Words in blue color shows the alignment with ground truth while words in red color mean wrong prediction part. Model trained on our dataset demonstrates more reasonable answers.

Algorithm 1: The Paradigm of DreamFrame

Require: In-context learned LLM G, Diffusion model ϵ_θ.
Input: movie description D, prompts P.
Output: Style descriptions S, Frame descriptions FD, Init style image set I, Overview O, Style embedding v, Visual frames VF, QA pairs Q.

$S, O \leftarrow G(P, D)$;
$FD, Q \leftarrow G(P, O)$ using story expanding;
for $i \leftarrow 1$ ***to*** 10 **do**
  $x_t \sim \mathcal{N}(0, I)$;
  for $t \leftarrow T$ ***to*** 1 **do**
    $x_{t-1} \leftarrow \epsilon_\theta(x_t, c_\theta(S), t)$;
  end
  Add x_0 to I;
end
for $j \leftarrow 1$ ***to*** $len(I)$ **do**
  for $t \leftarrow T$ ***to*** 1 **do**
    $v_t \leftarrow$ Optimized by LDM Loss with gt noise $\epsilon(I_j)$;
  end
end
for $k \leftarrow 1$ ***to*** $len(FD)$ **do**
  $FD_k \leftarrow$ Prefix with a style word + FD_k;
  $x_t \sim \mathcal{N}(0, I)$;
  for $t \leftarrow T$ ***to*** 1 **do**
    $x_{t-1} \leftarrow \epsilon_\theta(x_t, c_\theta\{v\}(FD_k), t)$;
  end
  Add x_0 to VF;
end

CHATGPT PROMPT FOR MOVIE PLOT GENERATION

You are a specialist in creating stories. Generate a story which theme is "[theme]". Requirements: First, generate the title, overview and "story tone and the style" of the story. I will use the story to generate images, so "Story tone and style" should describe the artistic style of the pictures, and should use only adjectives and short phrases, not sentences, with the most important phrases included in "(("))", and a single one-word style-keyword included in "{""}"(Note that you must provide the style-keyword like the example). Second, list all the locations that the story will take place in, at least 12 locations. Only common places are allowed and use at most 2 words for the location. Do not mention character names in the locations. Do not use location like "Mary's" or "Mary's house" to indicate someone's home. Third, generate all the characters' name, and use the name of a Western celebrity to describe them, with clothes color, in the format of "<xxx: a man/woman looks like xxx in xxx>". For clothes color, a simple one-word color is enough. If a character appears in the story for at least twice, put it in "Character" part. A main charater appears most frequently, and a supporting character appears in part of the story. Except for the characters listed in "Character" part, all the other characters and objects that will appear in the story should only appear once. For the new story, I need 1 or 2 main characters, and 0 or 1 supporting charater. Fourth, generate 11 substories based on the overview. "Scene" part should pick 1,2 or 3 locations in the "Locations" part, and the whole substory will only take place in these locations. "QA" part is a question-answer pair based on each substory. You can only ask one of the four types of questions"what", "why", "how" and "where" after each substory. The answer should be one or two sentences, no need to be too long. Below is an example:

Story Title: The Symphony of Love

Overview: 'The Symphony of Love' is a heartwarming tale set in the bustling city of New York, painted in a vibrant, contemporary style. The story revolves around Amelia, a talented violinist with a passion for music, and Ethan, a successful businessman with a hidden love for the arts. Their lives intersect as they navigate the complexities of love, ambition, and the pursuit of dreams.

Story tone and style: {Romantic}, ((Urban chic)), ((modern elegance, bohemian charm)), pastel palette, minimalist aesthetic, soft lighting, candid moments, architectural beauty, cityscape, dynamic, emotional, intimate, dreamy, nostalgic

Style-keyword: Romantic

Locations: ['Concert Hall', 'Office', 'Cafe', 'Art Gallery', 'Park', 'Restaurant', 'Apartment', 'Subway', 'Rooftop', 'Street', 'Lake', 'Classroom', 'Cinema']

Character:
Main characters:
<Amelia: a woman looks like Anne Hathaway in red>
<Ethan: a man looks like Chris Hemsworth in blue>

Supporting character:
<Oliver: a man looks like Robert Downey Jr. in brown>

Substories:

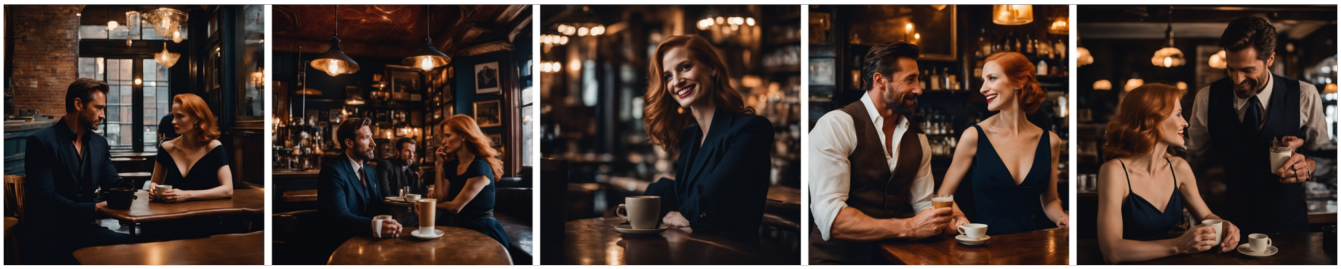
Substory1: The Melody of Chance
Scene: Subway, Street
Content: Amelia and Ethan have a chance encounter in the subway, and walk on the street.
QA: Question: Where did Amelia and Ethan meet each other? Answer: In the subway

Substory2: The Harmony of Friendship
Scene: Cafe
Content: Amelia and Ethan start to build a friendship over shared coffees and conversations at their favorite cafe.
QA: Question: How did Amelia and Ethan know each other? Answer: They meet each other in cafe, drinking coffee.

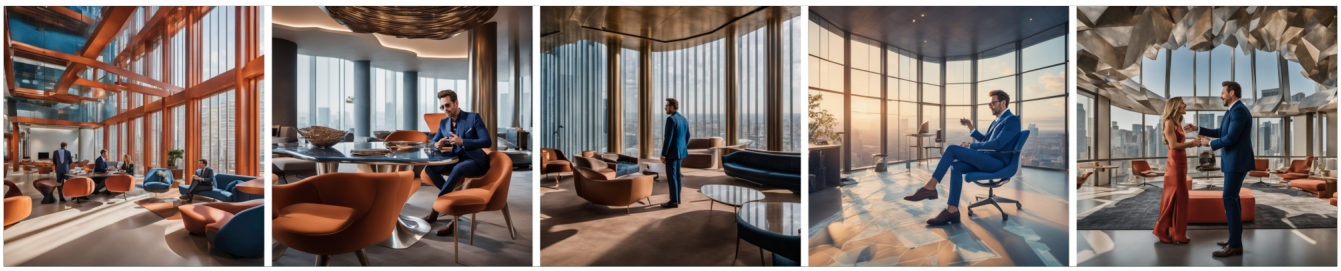
I didn't list all the substories, but you should generate the complete version.

Figure A: Prompt for Movie Plot Generation. Our prompt for movie plot generation comprises basic elements, including overview, characters, style and so on.

generate an image in Comedic style: In the newspaper office, a man looks like Hugh Laurie in black holds a portfolio labeled 'Wedding of the Century', while standing near a large glass window, looking thoughtful.	generate an image in Comedic style: In the newspaper office, a man looks like Hugh Laurie in black, holding his black pen and the portfolio, sitting on his desk, with a troubled expression on his face.	generate an image in Comedic style: In the newspaper office, a man looks like Hugh Laurie in black walks past a co-worker carrying a huge wedding cake model, with a baffled look on his face.	generate an image in Comedic style: In the newspaper office, a man looks like Hugh Laurie in black, standing near the entrance door, gazes now at the real wedding cake that just arrived, with a look of disbelief.	generate an image in Comedic style: In the newspaper office, a man looks like Hugh Laurie in black sits at his desk, head in hands, looking at the portfolio, looking bewildered.

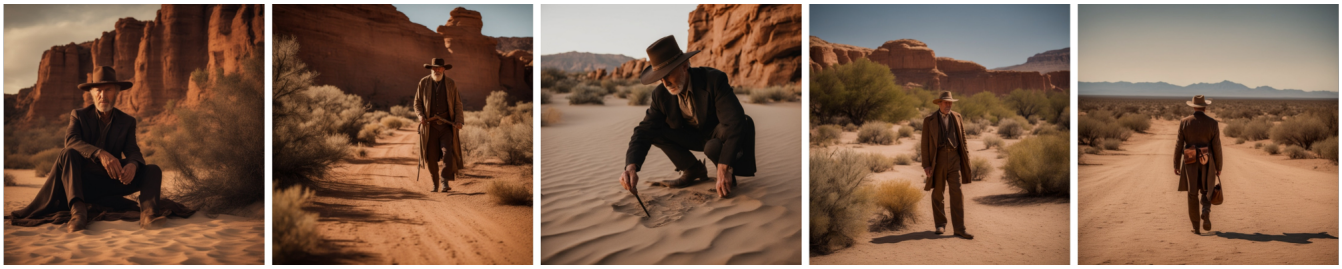


generate an image in tense style: In the coffee shop, a man looks like Hugh Jackman in black is still sitting across from a woman looks like Jessica Chastain in navy, looking at her, waiting for a response.	generate an image in tense style: In the coffee shop, a woman looks like Jessica Chastain in navy narrows her eyes and looks at a man looks like Hugh Jackman in black, thinking.	generate an image in tense style: In the coffee shop, a man looks like Hugh Jackman in black lifts his cup towards a woman looks like Jessica Chastain in navy for a toast.	generate an image in Comedic style: In the newspaper office, a man looks like Hugh Laurie in black, standing near the entrance door, gazes now at the real wedding cake that just arrived, with a look of disbelief.	generate an image in tense style: In the coffee shop, a woman looks like Jessica Chastain in navy lifts her cup towards a man looks like Hugh Jackman in black in response, with a smile.



generate an image in Dynamic style: In the modern office building, a man looks like Robert Downey Jr. in blue is sitting across from the investor, looking at her with hope, showing her his groundbreaking concept.	generate an image in Dynamic style: In the modern office building, a man looks like Robert Downey Jr. in blue, looking a bit sceptical, is sitting beside the metallic silver model of his concept on the table.	generate an image in Dynamic style: In the modern office building, a man looks like Robert Downey Jr. in blue is standing by the panoramic window, looking out, with a speculative expression.	generate an image in Dynamic style: In the modern office building, a man looks like Robert Downey Jr. in blue is sitting in a chair near the window, looking at a visible holographic screen with his idea blueprints, visibly happy.	generate an image in Dynamic style: In the modern office building, a man looks like Robert Downey Jr. in blue, with a huge smile on his face, is standing in the middle of the room, hugging the investor a woman looks like Jennifer Aniston in black, after sorting out their collaboration details.

Figure B: Examples of generated long video instruction data We use GPT-4 and guided text-to-image generation models to generate consistent key frames of move-level video with reasonable lines and corresponding question-answer pairs. These data are used to train multimodal large language models on video understanding.



generate an image in Gritty style: In the desert, a man looks like Clint Eastwood in black is sitting on the desert sand, his gaze fixed on the hidden faces in the outlaw camp, with a grave face.	generate an image in Gritty style: In the desert, Monroe is walking cautiously, approaching the outlaw camp while holding a small stick, his face serious.	generate an image in Gritty style: In the desert, a man looks like Clint Eastwood in black is squatting, using the stick to draw a map in the sand, with a determined expression.	generate an image in Gritty style: In the desert, Monroe is standing, scanning the camp once more, a mixture of anticipation and concern etched on his face.	generate an image in Gritty style: In the desert, Monroe is walking toward the camp, a saddlebag across his shoulder and eyes on the horizon, his face hardened.

generate an image in sad style: At home, a boy looks like Timothee Chalamet in black is standing in the living room, looking out of the window, with an anguished face.	generate an image in sad style: At home, a boy looks like Timothee Chalamet in black is sitting in his room, holding a map of the countryside, with a thoughtful face.	generate an image in sad style: In the countryside, a boy looks like Timothee Chalamet in black is walking across an open field, with a relieved face.	generate an image in sad style: In the countryside, a boy looks like Timothee Chalamet in black is sitting against a haystack, looking at the setting sun, with a peaceful face.	generate an image in sad style: In the countryside, a boy looks like Timothee Chalamet in black is standing near a stream, watching the water flow, with a calm face.

generate an image in Thrilling style: In the lab, a woman looks like Nicole Kidman in white is walking towards a telephone on her desk, a look of resolve returning to her face.	generate an image in Thrilling style: In the lab, a woman looks like Nicole Kidman in white is standing by the window, looking out, a worried look clouding her face.	generate an image in Thrilling style: In the lab, a woman looks like Nicole Kidman in white is sitting down, looking at the countdown clock on her desk, her eyes filled with determination.	generate an image in Thrilling style: In the lab, a woman looks like Nicole Kidman in white is standing, holding her lab coat, looking ready to fight, a determined expression hardening her face.	generate an image in Thrilling style: In the lab, a woman looks like Nicole Kidman in white is walking out the door, holding her file of evidence, her face full of determination for the task ahead.

Figure C: Examples of generated long video instruction data We use GPT-4 and guided text-to-image generation models to generate consistent key frames of move-level video with reasonable lines and corresponding question-answer pairs. These data are used to train multimodal large language models on video understanding.

Can you provide a brief overview of the movie in just few sentences?

Reference: 'The Pursuit of Happyness' is an inspiring drama based on the true story of Chris Gardner, a struggling salesman who becomes homeless with his young son. Faced with financial instability and life's adversities, Gardner perseveres through immense challenges to secure a better future for himself and his son. The film explores themes of resilience, determination, and the American dream, as Gardner embarks on a grueling journey to change their lives, ultimately landing a coveted position as a stockbroker after completing a highly competitive internship.

Ours: 'The Pursuit of Happyness' follows the life of Chris Gardner, a struggling stockbroker in San Francisco, who is fired from his job due to his family problems and financial struggles. Despite his setback, Chris perseveres by taking a job as a broker, where he discovers a passion for the financial industry and begins to build his own wealth. His life becomes more complicated when he becomes the target of a con artist named Jay Twistle, who convinces him to invest in a scanner, which turns out to be a time machine. The story culminates in a series of events that test Chris's sanity, and he is forced to confront the reality of his situation, leading him to confront the pursuit of happiness and the importance of living a simple life

LLaMA-VID: In \"Riding the Bus,\" Chris Gardner, a man with big dreams, is forced to take a lower-paid job to support his family after his father's arrest. He is determined to make it big in the financial world and eventually gets a job at a brokerage firm. Along the way, he faces personal challenges like a broken leg and a car accident that forces him to walk to work, but he remains determined and eventually succeeds in his goal. The film ends with Chris and his son, Christopher, embracing the happier aspects of life, as they ride the bus together.

What leads Chris Gardner and his son to become homeless?

Reference:Chris Gardner and his son become homeless due to a combination of financial difficulties stemming from unsold inventory of bone density scanners, unpaid taxes, and increasing living expenses. Gardner's struggle is compounded when his wife leaves him, and he is unable to afford their apartment rent, leading to their eviction. This series of events forces Chris and his son to seek temporary shelter wherever possible, including in homeless shelters, public restrooms, and subway stations.

Ours: Chris Gardner and his son become homeless due to a series of events. First, they have to leave the motel where they were living. Then Chris gets fired from his job at the brokerage firm due to the high cost of the scanner he bought. Next, his wife Linda leaves him, causing them to lose the motel's rent. Finally, their car breaks down, leaving them stranded in the parking lot. All these events lead them to live in a homemade camp in a warehouse, where they are forced to scavenge for food and struggle to survive in a harsh environment.

LLaMA-VID: Chris Gardner and his son become homeless after being evicted from their motel.

Can you outline the chronological sequence of events in the movie?

Reference:The film starts with Chris Gardner's struggle to sell medical equipment while grappling with financial instability. His wife leaves due to the mounting pressures, making Gardner a single parent. They are evicted from their apartment, leading to a life of homelessness. Gardner learns of a competitive internship at a brokerage firm, seeing it as an opportunity for a better life. Despite the challenges of homelessness, Gardner commits to the unpaid internship, balancing fatherhood, financial difficulties, and the rigorous demands of the job. His tenacity is tested through various trials, including spending nights in homeless shelters and public restrooms. The movie culminates in the emotional moment when Gardner is informed that he has secured a position as a stockbroker, a victory achieved through his resilience and hard work. The film concludes with Gardner walking hand in hand with his son, stepping into a future filled with hope and possibilities, reflecting on their journey from despair to happiness.

Ours: The movie follows the life of Chris Gardner, a young man who is struggling to escape poverty and gain wealth through his work at a brokerage firm. He first encounters Linda and her son, and later becomes an intern at a brokerage firm where he is fired. Chris is arrested for running away from a hitman, and he is forced to take his son, Tim, to a motel. After a series of failed attempts, including a failed interview with a football team, Chris gets an internship with an investment firm where he is fired for being too unreliable. In a desperate move, Chris steals a scanner from a pawn shop and sells it to a medical equipment company. He then becomes an insanely successful broker, selling the scanner for a fortune and buying a one-way ticket to New York. Chris's wealthy past, his struggles with his son's upbringing, and his own inner demons are explored through his interactions with his son and his brokerage colleagues.

LLaMA-VID:
1. Chris Gardner starts a new life in San Francisco after discovering his wife Linda is pregnant.
2. He takes up a job at a daycare and struggles to balance his responsibilities with his son.
3. Chris's father, a homeless man, comes to San Francisco, and Chris's life is further disrupted when he has to take care of his son after Linda's sudden death.
4. Chris is arrested for failing to pay parking tickets and is forced to spend a night in jail.
5. Chris's father is homeless, and he is arrested for loitering.
6. Chris is hired by a brokerage firm and begins an internship, where he is ridiculed by his peers.
7. Chris is hired by a rival firm, where he is ridiculed but determined to make a name for himself.
8. Chris's father is arrested again, and Chris is forced to take care of his son for the second time.
9. Chris's father is arrested again, and Chris is forced to take care of his son again.
10. Chris's father is arrested again, and Chris is forced to take care of his son again.
11. Chris is arrested for the third time, and this time he is forced to take care of his son again.
12. Chris is hired by a rival brokerage firm, where he is ridiculed again but determined to make a name for himself.
13. Chris's father is arrested again, and Chris is forced to take care of his son again.

Figure D: Video understanding results of baseline model and ours. Our model demonstrates a more reasonable and detailed answer in the video understanding.

REFERENCES

[1] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. 2023. InstructBLIP: Towards General-purpose Vision-Language Models with Instruction Tuning. *arXiv:2305.06500* (2023).

[2] Yuxin Fang, Wen Wang, Binhui Xie, Quan Sun, Ledell Wu, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. 2023. Eva: Exploring the limits of masked visual representation learning at scale. In *CVPR*.