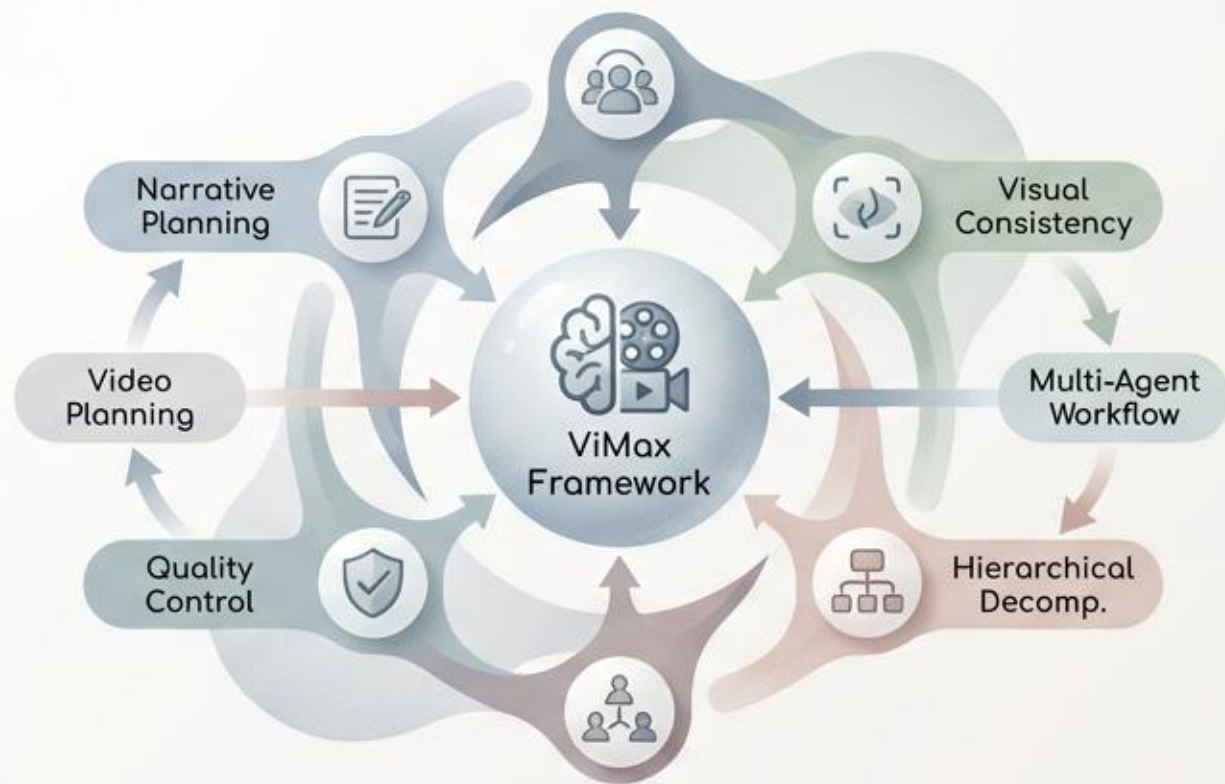


ViMax: Agentic Video Generation



Lingxuan Huang, Sizhe He, Hengji Zhou, Yan Huang, Si Wu, Lianghao Xia, Chao Huang.
The University of Hong Kong and South China University of Technology.

Research Background and Problem



Challenges in Long-Form Video Generation:
The ViMax framework addresses significant challenges in long-form video generation, focusing on narrative planning, visual consistency, and integration of narrative and visual elements.



Limitations of Current Models:
Current models are limited to short-form content and face issues like 'catastrophic semantic forgetting,' semantic drift, independent shot generation, and limited narrative planning.



Necessitating New Frameworks:
These limitations impede the production of coherent long-form content, necessitating new frameworks like ViMax that can manage these complexities.

CHALLENGES

Narrative Planning, Visual Consistency, Integration



LIMITATIONS

- Short-Form Only
- Semantic Forgetting & Drift
- Independent Shots
- Limited Planning



SOLUTION: VIMAX FRAMEWORK

- Coherent Long-Form Content
- Manage Complexities
- Multi-Agent System



Framework Overview: ViMax



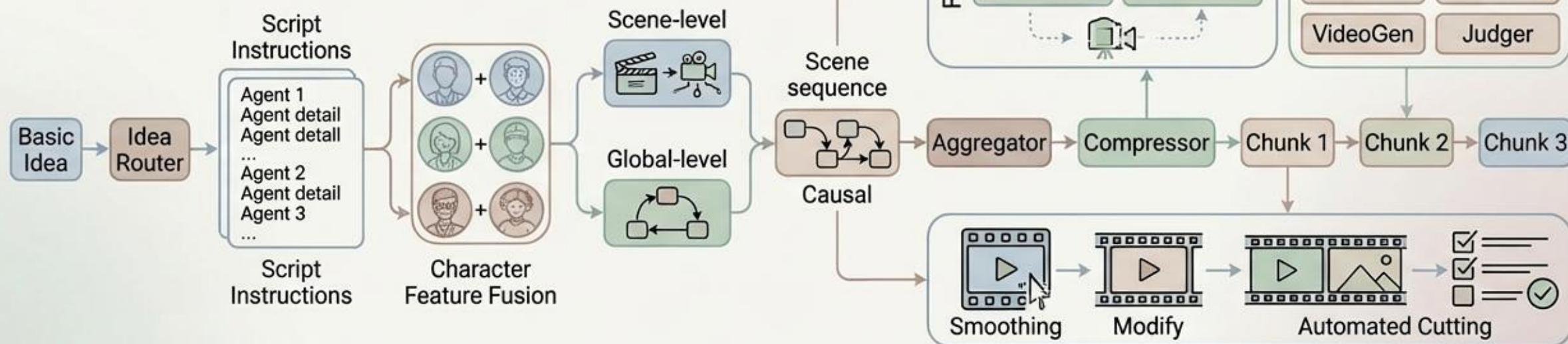
ViMax is an agentic video generation framework employing a multi-agent system for producing long-form videos. The system coordinates agents specializing in screenwriting, shot planning, character styling, video generation, and quality control.



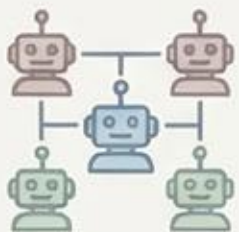
Key methodologies include hierarchical story decomposition, retrieval-augmented generation (RAG), graph-based visual consistency, and VLM-guided quality control.

$$\mathcal{S} = \text{ScreenWriting}(\mathbb{Z}; P_{C(\mathbb{Z})})$$

where \mathbb{Z} is the user narrative input leading to screenplay \mathcal{S} .



Key Components of ViMax



Multi-Agent
Workflow

The framework's components include multi-agent workflow for collaborative generation processes; hierarchical story decomposition for manageable narrative planning; retrieval-augmented generation enhancing local narrative descriptions; and visual consistency mechanisms using graph-based dependency tracking. Additionally, VLM-guided quality control ensures visual fidelity. The framework employs a two-step text-image-video generation: $\{ K1, K2, ..., Km \} = \text{ImageGen}(\{ S \})$ and $\{ V1, V2, ..., Vn \} = \text{VideoGen}(\{ S \}, \{ K \})$.



Hierarchical Story
Decomposition



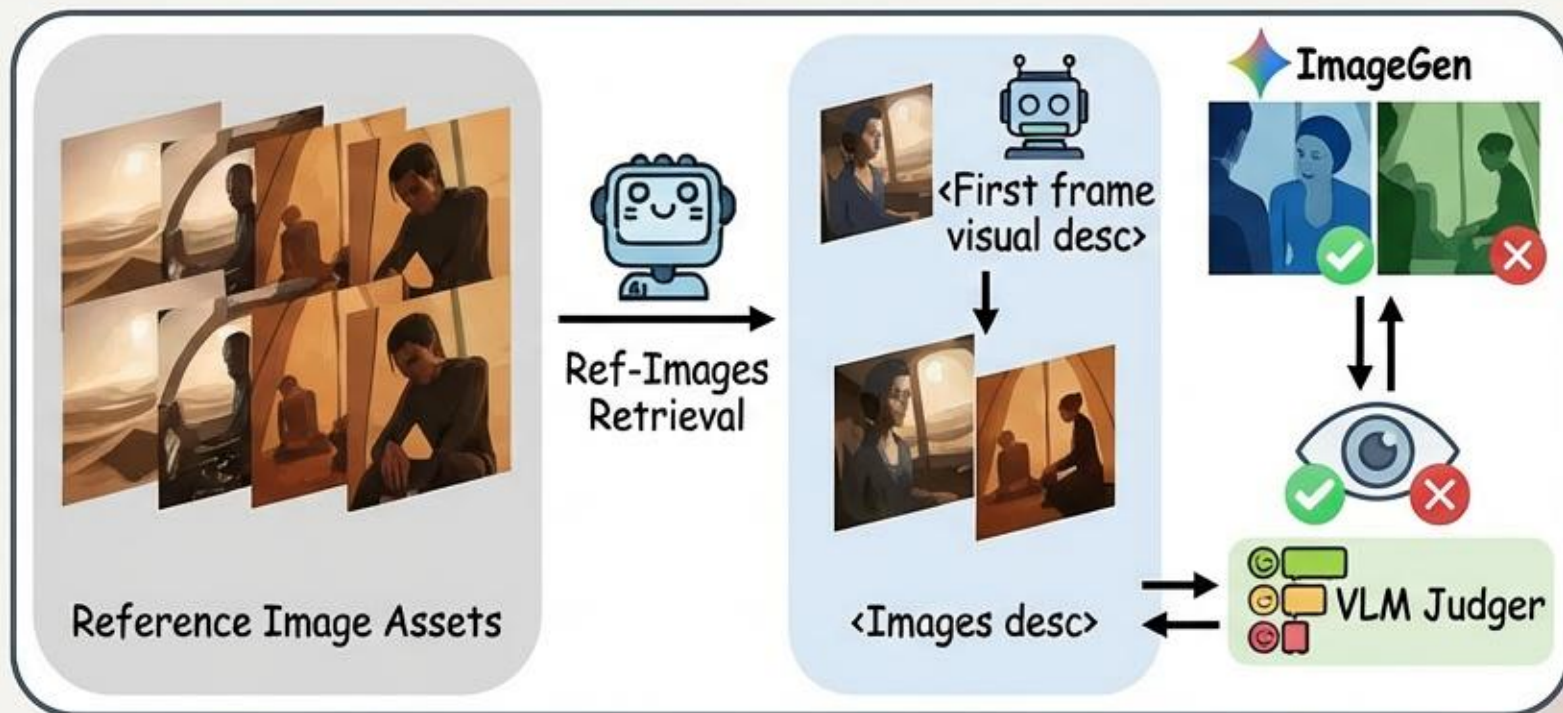
RAG - Narrative
Descriptions



Visual Consistency -
Graph-based



VLM-Guided
Quality Control



VLM-based Quality Control Workflow

Experimental Results



ViMax is evaluated on datasets designed for long-form narratives, focusing on maintaining narrative coherence and visual consistency.



Evaluation metrics include thematic coherence and visual fidelity.



The framework demonstrated significant improvement over baseline models in maintaining story coherence across extended.



The framework demonstrated significant improvement over baseline models in maintaining story coherence across extended sequences.

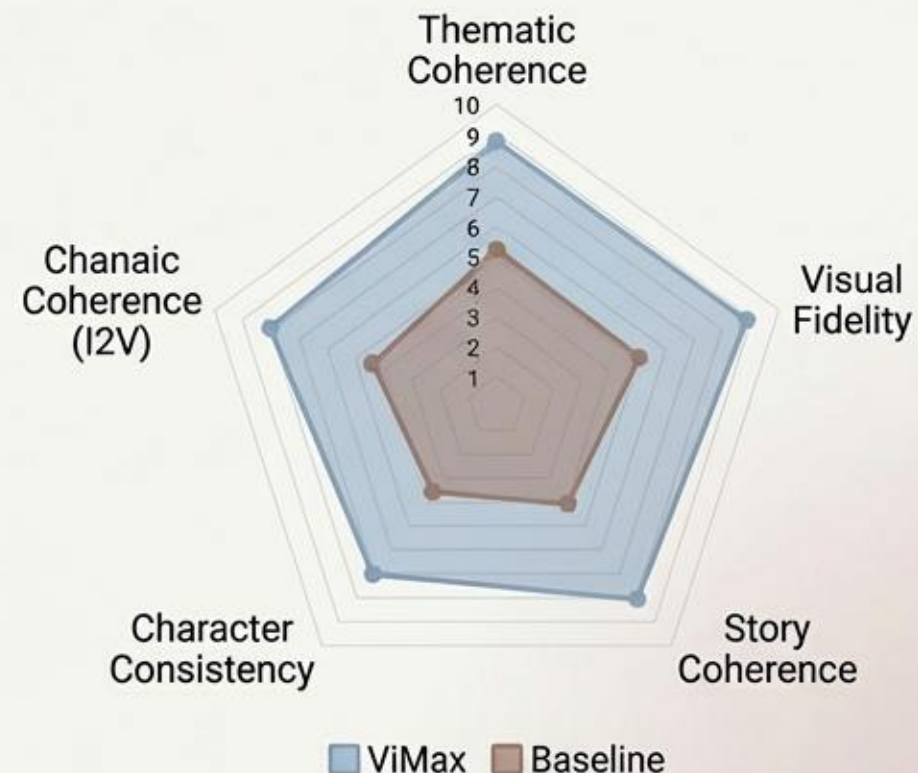


Figure 5: Case study: ViMax-generated keyframes.

Conclusion and Contributions



Addresses critical gaps in long-form video generation.



Ensures narrative and visual consistency using a multi-agent approach.



Key contributions: Hierarchical story decomposition and retrieval-augmented generation.



Effectiveness in producing coherent long-form video content, enhancing future methodologies.

VLM-guided Selection Example



VLM Judger System Prompt:

"You are a professional visual assessment expert. Given:

(1) reference images with short text descriptions,

(2) candidate images,

(3) a target text description between <TARGET_DESCRIPTION_START> and <TARGET_DESCRIPTION_END>,"

select the best candidate image based on: 1. Character Consistency: match of gender, ethnicity, age, facial features, body shape, outfit, hairstyle to the reference. 2. Spatial Consistency: match of relative positions, scene layout, and perspective to the reference. 3. Description Accuracy: how well the image matches the target text description (which specifies the desired final image, not an edit instruction).



Candidate 1



Candidate 2



Selected Image



VLM response:

```
{
  "best_image_index": 2,
  "reason": "All three candidate images closely match the references and description, but Candidate Image 2 best captures the specified actions.\n\n**Candidate Image C:**\n- Character Consistency: 9/10. The boy and dog closely match the references.\n- Spatial Consistency: 10/10. The backyard layout with the boy on the left and dog on the right is correct.\n- Description Accuracy: 8/10. The dog is standing rather than crouched in anticipation.\n\n**Candidate Image 1:** This makes it the best overall choice."
```