AutoMV: An Autonomous Agent Framework for Real Estate Marketing Video Generation

Kuizong Wu, Shaozu Yuan, Chang Shen, Long Xu, Meng Chen[*]

Yep AI, Melbourne, Australia
{vincent.wu, shaozu.yuan, chang.shen, neo.xu}@yepai.io, chenmengdx@gmail.com

Abstract

In this paper, we introduce AutoMV, an autonomous agent framework designed for generating real estate marketing videos. The framework integrates a diverse set of existing models into a tool library, allowing the agent to intelligently select and execute the appropriate tools. Given property images and text, the agent decomposes the task into manageable subtasks, generating storyline directives and corresponding camera movement trajectories to guide the video production process. By automatically applying video synthesis techniques and incorporating multimedia elements such as subtitles and background music, the agent transforms static real estate images into dynamic, visually appealing videos, thereby optimizing their impact for digital marketing purposes.

Introduction

Visual media is crucial for real estate digital marketing, offering detailed product presentations and boosting potential buyer engagement. However, most real estate platforms still rely heavily on static images, which is insufficient to deliver the spatial awareness and immersive experience videos can provide. This gap primarily stems from the high costs and complexity of producing video content, which are not feasible for individual homeowners and smaller real estate firms. Attempts to address this, such as creating slideshows with transitional effects to simulate videos (Hua, Lu, and Zhang 2006; Luo et al. 2012), reduce production challenges but fail to achieve the same coherence and viewer engagement, often resulting in viewer fatigue due to repetitive transitions.

Recent advancements (Ho et al. 2022b,a; Singer et al. 2022; Blattmann et al. 2023; Chen et al. 2023a; Girdhar et al. 2023) demonstrate diffusion models' impressive capabilities in video generation, with models capable of producing extended videos based on conditional imagery. Some of these models (Xing et al. 2023; Zhang et al. 2023b; Chen et al. 2023b; Ren et al. 2024; Zhang et al. 2024; Sharma et al. 2024) focus on enhancing the predictability of video outputs driven by text. For instance, OpenAI's Sora (Brooks et al. 2024) is capable of producing videos exceeding a minute in duration guided by textual prompts. However, text-prompt-based generation models often struggle to capture the intricate details within prompts, making it challenging to ensure the stability and predictability of the generated video (Fan et al. 2024). Additionally, this approach still depends on the user's skill in crafting effective prompts and their expertise in video production, resulting in inconsistent video quality. Furthermore, fine-grained motion control based on storyline directives is essential but remains largely unaddressed.

To address these challenges, we introduce **AutoMV**, an innovative marketing video generation system powered by a multimodal large language model (MLLM) agent (Ma, Zhang, and Zhao 2024; Wang et al. 2023a). AutoMV autonomously transforms static real estate images into dynamic, engaging videos by generating narrative directives and corresponding camera movement trajectories, effectively constructing a dynamic storyboard. Leveraging advanced vision-language models (Liu et al. 2023a; Team et al. 2023; Anthropic 2024), the agent analyzes images to identify salient features, optimize camera movements (e.g., panning, zooming, transitions), and localize objects for precise framing and cropping. Subsequently, a video generation diffusion model synthesizes videos that follow the specified camera movements, resulting in a visually dynamic and compelling presentation of the real estate properties. Furthermore, AutoMV incorporates multimedia elements such as subtitles and background music, chosen based on contextual relevance and aesthetic harmony. This comprehensive approach significantly enhances the presentation and marketing impact of real estate listings, offering a cost-effective and highly automated solution for individual homeowners and smaller real estate agencies.

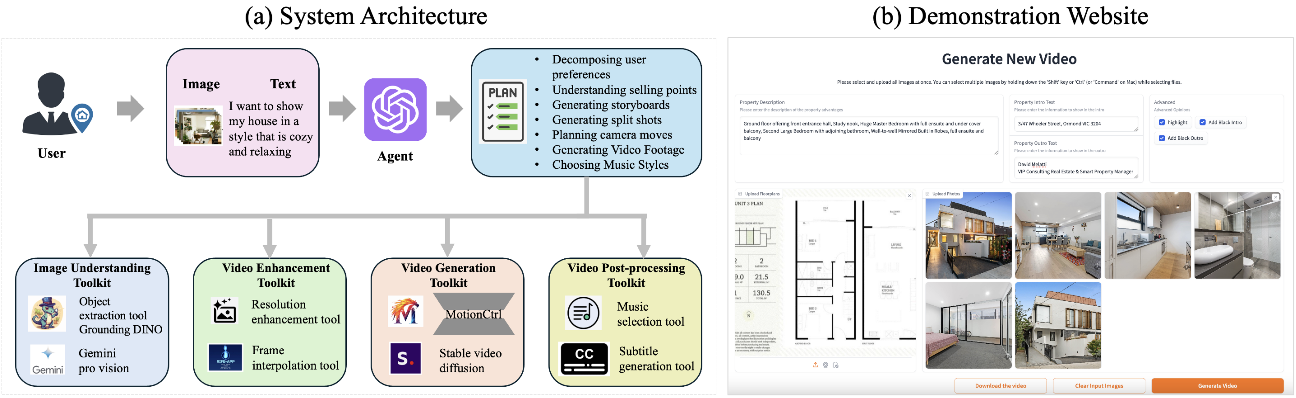
System Architecture

Figure 1 illustrates an overview of the AutoMV framework. At the core of the system is the MLLM agent, which functions as the "brain" to autonomously orchestrate the entire workflow. It decomposes tasks and formulates the solution using specialized tools to execute key processes, including image understanding, video enhancement, video generation, and post-processing of video.

Agent Planing

After receiving the image and property description inputs, the MLLM agent autonomously decomposes the problem into subtasks. Utilizing GPT-4V's robust scene recognition

[*] Corresponding author



(a) System Architecture

(b) Demonstration Website

Generate New Video

Please select and upload all images at once. You can select multiple images by holding down the 'Shift' key or 'Ctrl' (or 'Command' on Mac) while selecting files.

Property Description
Please enter the description of the property advantages
Ground floor offering front entrance hall, Study nook, Huge Master Bedroom with full ensuite and under cover balcony, Second Large Bedroom with adjoining bathroom, Wall-to-wall Mirrored Built in Robes, full ensuite and balcony

Property Intro Text
Please enter the information to show in the intro
3/47 Wheeler Street, Ormond VIC 3204

Property Outro Text
Please enter the information to show in the outro
David Melatti
VIP Consulting Real Estate & Smart Property Manager

Advanced
Advanced Opinions
☑ highlight ☑ Add Black Intro
☑ Add Black Outro

Figure 1: System architecture and demo website (screenshot)

and inference capabilities (OpenAI 2023), it performs comprehensive image analysis to extract salient features and unique selling points (Yang et al. 2023; Zhou et al. 2023). By interpreting user preferences and marketing objectives, the agent generates customized storylines, plans camera movements (e.g., zoom in, pan up), and transitions to align with the desired narrative. This includes designing paths to highlight specific property features, effectively showcasing the property's appeal. The agent then executes each subtask by calling external tools, ultimately solving the problem.

Object Extraction

The object extraction tool within the image understanding toolkit enables the agent to process images and associated text, identifying objects and generating close-up shots of these elements. For enhanced object detection, the agent utilizes Grounding DINO (Liu et al. 2023b), an open-set detector that integrates the Transformer-based DINO detector (Zhang et al. 2023a) with grounded pre-training. Given that the objects identified often stem from an infinitely large and unpredictable input space (Li et al. 2023), the zero-shot detection capability of this module is crucial for the agent's adaptability and efficacy.

Resolution Promotion

The video diffusion model requires fixed-resolution input images, and simple resizing can lead to detail loss and instability. To mitigate this, the MLLM agent assesses image resolution to decide on using a resolution enhancement tool before video generation. Specifically, the agent employs the SWINIR super-resolution model (Liang et al. 2021) to upscale images, preserving original details and minimizing quality degradation during diffusion. This method balances inference speed and algorithmic efficiency, providing the agent with improved images for subsequent processing.

Video Generation

For the video generation process, the MLLM agent leverages the advanced video diffusion model, MotionCtrl (Wang et al. 2023b), due to its ability to precisely control camera trajectories in accordance with the project's specifications. This model integrates stable video diffusion technology with a sophisticated camera motion control module. The agent guides virtual camera's movements according to the combination of visual input and its own planning from the image understanding module, ensuring adherence to predetermined trajectories within spatial coordinates to effectively showcase property features.

Video Post-Processing

To enhance the final video output, the MLLM agent employs a video post-processing module. Given that MotionCtrl (Wang et al. 2023b) produces video clips at 14 frames per second (fps), the agent uses RIFE (Huang et al. 2022), an optical flow-based real-time interpolation method, to increase the frame rate and improve smoothness. Furthermore, the agent customizes the video by adding subtitles, selecting appropriate background music based on user preferences and image characteristics, and incorporating informative text overlays that display property details or floor plans at the beginning and end of the video. This ensures the generated videos are not only visually appealing but also informative and engaging.

Demonstration

We created a demo website[1] to help users easily produce real estate showcase videos. Users can upload images, select features, and add text, all without dealing with technical complexities. For more details, please watch the video in the supplementary materials.

Conclusion

This paper presents a real estate video generation agent that transforms static real estate images into dynamic videos by leveraging vision-language models and diffusion techniques. The agent autonomously analyzes user input, decomposes the task, selects appropriate tools, and executes the most suitable processes, offering a cohesive framework for marketing video creation. Future work will focus on expanding the tool library to augment the system's capabilities.

[1] Demo website: https://automv.yepai.com.au. More generated video showcases here: https://youtu.be/aChD6FHVFK4

References

Anthropic, A. 2024. The claude 3 model family: Opus, sonnet, haiku. *Claude-3 Model Card*, 1.

Blattmann, A.; Dockhorn, T.; Kulal, S.; Mendelevitch, D.; Kilian, M.; Lorenz, D.; Levi, Y.; English, Z.; Voleti, V.; Letts, A.; Jampani, V.; and Rombach, R. 2023. Stable Video Diffusion: Scaling Latent Video Diffusion Models to Large Datasets. arXiv:2311.15127.

Brooks, T.; Peebles, B.; Holmes, C.; DePue, W.; Guo, Y.; Jing, L.; Schnurr, D.; Taylor, J.; Luhman, T.; Luhman, E.; Ng, C.; Wang, R.; and Ramesh, A. 2024. Video generation models as world simulators.

Chen, H.; Xia, M.; He, Y.; Zhang, Y.; Cun, X.; Yang, S.; Xing, J.; Liu, Y.; Chen, Q.; Wang, X.; Weng, C.; and Shan, Y. 2023a. VideoCrafter1: Open Diffusion Models for High-Quality Video Generation. arXiv:2310.19512.

Chen, X.; Wang, Y.; Zhang, L.; Zhuang, S.; Ma, X.; Yu, J.; Wang, Y.; Lin, D.; Qiao, Y.; and Liu, Z. 2023b. SEINE: Short-to-Long Video Diffusion Model for Generative Transition and Prediction. arXiv:2310.20700.

Fan, F.; Luo, C.; Gao, W.; and Zhan, J. 2024. AIGCBench: Comprehensive Evaluation of Image-to-Video Content Generated by AI. arXiv:2401.01651.

Girdhar, R.; Singh, M.; Brown, A.; Duval, Q.; Azadi, S.; Rambhatla, S. S.; Shah, A.; Yin, X.; Parikh, D.; and Misra, I. 2023. Emu Video: Factorizing Text-to-Video Generation by Explicit Image Conditioning. arXiv:2311.10709.

Ho, J.; Chan, W.; Saharia, C.; Whang, J.; Gao, R.; Gritsenko, A.; Kingma, D. P.; Poole, B.; Norouzi, M.; Fleet, D. J.; and Salimans, T. 2022a. Imagen Video: High Definition Video Generation with Diffusion Models. arXiv:2210.02303.

Ho, J.; Salimans, T.; Gritsenko, A.; Chan, W.; Norouzi, M.; and Fleet, D. J. 2022b. Video diffusion models. *Advances in Neural Information Processing Systems*, 35: 8633–8646.

Hua, X.-S.; Lu, L.; and Zhang, H.-J. 2006. Photo2Video—A System for Automatically Converting Photographic Series Into Video. *IEEE Transactions on Circuits and Systems for Video Technology*, 16(7): 803–819.

Huang, Z.; Zhang, T.; Heng, W.; Shi, B.; and Zhou, S. 2022. Real-Time Intermediate Flow Estimation for Video Frame Interpolation. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

Li, J.; Xie, C.; Wu, X.; Wang, B.; and Leng, D. 2023. What Makes Good Open-Vocabulary Detector: A Disassembling Perspective. arXiv:2309.00227.

Liang, J.; Cao, J.; Sun, G.; Zhang, K.; Van Gool, L.; and Timofte, R. 2021. SwinIR: Image Restoration Using Swin Transformer. *arXiv preprint arXiv:2108.10257*.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023a. Visual Instruction Tuning. arXiv:2304.08485.

Liu, S.; Zeng, Z.; Ren, T.; Li, F.; Zhang, H.; Yang, J.; Li, C.; Yang, J.; Su, H.; Zhu, J.; et al. 2023b. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*.

Luo, S.-J.; Shen, I.-C.; Hsu, H.; and Chen, B.-Y. 2012. Attention-oriented photo slideshow. In *SIGGRAPH Asia 2012 Posters*, 1–1.

Ma, X.; Zhang, Z.; and Zhao, H. 2024. Comprehensive Cognitive LLM Agent for Smartphone GUI Automation. *arXiv preprint arXiv:2402.11941*.

OpenAI. 2023. GPT-4V(ision) System Card.

Ren, W.; Yang, H.; Zhang, G.; Wei, C.; Du, X.; Huang, S.; and Chen, W. 2024. ConsistI2V: Enhancing Visual Consistency for Image-to-Video Generation. *arXiv preprint arXiv:2402.04324*.

Sharma, A.; Yu, A.; Razavi, A.; Toor, A.; Pierson, A.; Gupta, A.; Waters, A.; van den Oord, A.; and et al. 2024. VEO: Visual Exploration and Optimization. https://deepmind.google/technologies/veo/. Accessed: 2024-06-18.

Singer, U.; Polyak, A.; Hayes, T.; Yin, X.; An, J.; Zhang, S.; Hu, Q.; Yang, H.; Ashual, O.; Gafni, O.; et al. 2022. Make-A-Video: Text-to-Video Generation without Text-Video Data. In *The Eleventh International Conference on Learning Representations*.

Team, G.; Anil, R.; Borgeaud, S.; Wu, Y.; Alayrac, J.-B.; Yu, J.; Soricut, R.; Schalkwyk, J.; Dai, A. M.; Hauth, A.; et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Wang, Z.; Cai, S.; Liu, A.; Ma, X.; and Liang, Y. 2023a. Describe, Explain, Plan and Select: Interactive Planning with Large Language Models Enables Open-World Multi-Task Agents. *arXiv preprint arXiv:2302.01560*.

Wang, Z.; Yuan, Z.; Wang, X.; Chen, T.; Xia, M.; Luo, P.; and Shan, Y. 2023b. MotionCtrl: A Unified and Flexible Motion Controller for Video Generation. In *arXiv preprint arXiv:2312.03641*.

Xing, J.; Xia, M.; Zhang, Y.; Chen, H.; Yu, W.; Liu, H.; Wang, X.; Wong, T.-T.; and Shan, Y. 2023. DynamiCrafter: Animating Open-domain Images with Video Diffusion Priors. arXiv:2310.12190.

Yang, Z.; Li, L.; Lin, K.; Wang, J.; Lin, C.-C.; Liu, Z.; and Wang, L. 2023. The Dawn of LMMs: Preliminary Explorations with GPT-4V(ision). arXiv:2309.17421.

Zhang, H.; Li, F.; Liu, S.; Zhang, L.; Su, H.; Zhu, J.; Ni, L.; and Shum, H.-Y. 2023a. DINO: DETR with Improved DeNoising Anchor Boxes for End-to-End Object Detection. In *The Eleventh International Conference on Learning Representations*.

Zhang, S.; Wang, J.; Zhang, Y.; Zhao, K.; Yuan, H.; Qin, Z.; Wang, X.; Zhao, D.; and Zhou, J. 2023b. I2VGen-XL: High-Quality Image-to-Video Synthesis via Cascaded Diffusion Models. arXiv:2311.04145.

Zhang, Y.; Xing, Z.; Zeng, Y.; Fang, Y.; and Chen, K. 2024. Pia: Your personalized image animator via plug-and-play modules in text-to-image models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7747–7756.

Zhou, P.; Cao, M.; Huang, Y.-L.; Ye, Q.; Zhang, P.; Liu, J.; Xie, Y.; Hua, Y.; and Kim, J. 2023. Exploring Recommendation Capabilities of GPT-4V(ision): A Preliminary Case Study. arXiv:2311.04199.