Scaling Vision–Language Models for Pharmaceutical Long-Form Video Reasoning on Industrial GenAI Platform

Suyash Mishra[a], Qiang Li[b], Satyanarayan Pati[c], Srikanth Patil[c], Narendra Baddu [c]
[a]Roche, [b]Accenture, [c]Involead,
Correspondence: suyash.mishra@roche.com, qiang.i.li@accenture.com satyanarayan.pati@involead.com
srikanth.patil@involead.com, baddu.narendra@involead.com

Abstract

Vision–Language Models (VLMs) have shown strong performance on multimodal reasoning tasks, yet most evaluations focus on short videos and assume unconstrained computational resources. In industrial settings such as pharmaceutical content understanding, practitioners must process long-form videos under strict GPU, latency, and cost constraints, where many existing approaches fail to scale. In this work, we present an industrial GenAI framework that processes over 200,000 PDFs, 25,326 videos across eight formats (e.g., MP4, M4V, etc.), and 888 multilingual audio files in more than 20 languages. Our study makes three contributions: (i) an industrial large-scale architecture for multimodal reasoning in pharmaceutical domains; (ii) empirical analysis of over 40 VLMs on two leading benchmarks (Video-MME and MMBench) and proprietary dataset of 25,326 videos across 14 disease areas; and (iii) four findings relevant to long-form video reasoning: the role of multimodality, attention mechanism trade-offs, temporal reasoning limits, and challenges of video splitting under GPU constraints. Results show 3–8× efficiency gains with SDPA attention on commodity GPUs, multimodality improving up to 8/12 task domains (especially length-dependent tasks), and clear bottlenecks in temporal alignment and keyframe detection across open- and closed-source VLMs. Rather than proposing a new "A+B" model, this paper characterizes practical limits, trade-offs, and failure patterns of current VLMs under realistic deployment constraints, and provide actionable guidance for both researchers and practitioners designing scalable multimodal systems for long-form video understanding in industrial domains.

* Patent application to EPO: 25175653.2

1 Introduction

Large Language Models (LLMs) such as GPT-4 (Ouyang et al., 2022) and LaMDA (Thoppilan et al., 2022) have significantly improved access to complex information in domains including healthcare and public services (H&PS) (Li et al., 2025a; Zhang et al., 2026; Mishra et al., 2026; Ozmermer and Li, 2023). While these models excel at text-based reasoning, industrial use cases increasingly involve multimodal content spanning tables, graphs, charts, text, video and audio (Fu et al., 2024a). In pharmaceutical industry, such assets include clinical trial recordings, conference lectures, promotional materials, and multilingual patient educational videos (Zhang et al., 2023). Manual perception of these resources is inconsistent, costly, and infeasible at scale, particularly given compliance & computation constraints.

Practical use cases including chat agents (OpenAI, 2024), briefing agencies (OpenAI, 2023), document searching (Lewis et al., 2021), text/video summarization (Li et al., 2025b), and document quality checks (Yang et al., 2024) deliver transformative impacts on industry.

Current research has emphasized novel architectures or benchmark leaderboards (Singhal et al., 2022), but less attention has been given to how existing VLMs scale under practical GPU budgets, long-video scenarios, and compliance workflows. Challenges remain in producing reliable VLMs with growing amount of video data, reducing hallucinations and memory cost, improving quality of long video content reasoning, and addressing computational bottlenecks (Qu et al., 2025). Industrializing VLMs for large-scale data especially using closed-source models on proprietary data is also becoming increasingly urgent.

This work centers on a core **Research Question (RQ)**: how to scale VLM-based multimodal reasoning on long-form pharmaceutical videos under realistic industrial GPU constraints. To address this, we open-source a large-scale GenAI platform architecture designed for natural language search for (H&PS) users. The platform ingested over 200,000

Proceedings of the 4th Workshop on Advances in Language and Vision Research (ALVR), pages 85–104

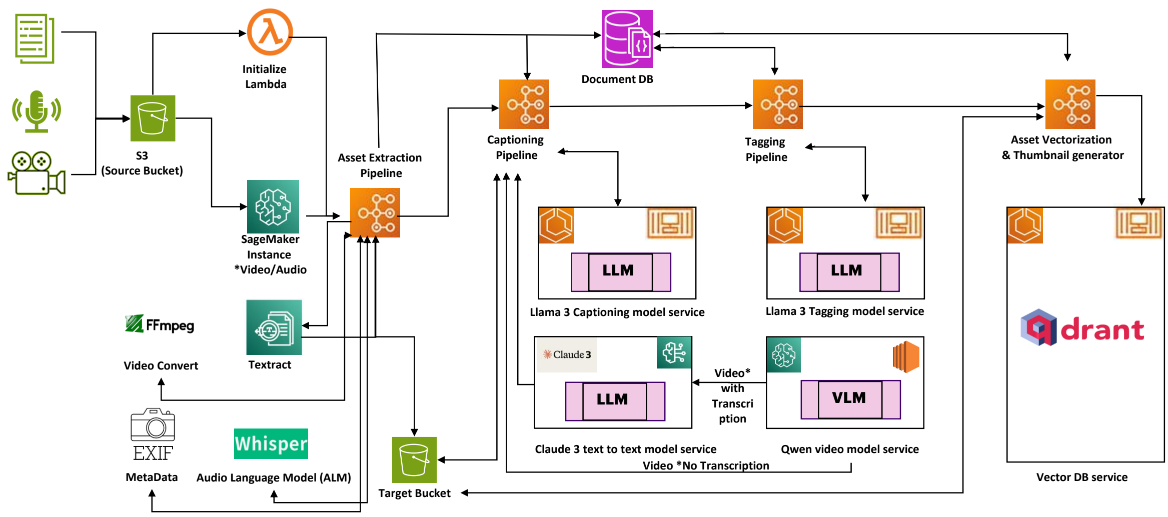


Figure 1: System architecture of our GenAI platform for Natural Language (NL) search integrating LLMs, ALMs, and VLMs. The platform processed 25,326 videos, 888 audios covering > 20 languages.

PDFs, 25,326 videos across 8 formats (e.g., MP4, M4V, MSVideo, etc.), and 888 audio files spanning 20+ languages. We evaluate more than 40 VLMs using leading benchmarks (Video-MME, MMBench) and our proprietary dataset covering 14 disease areas. Our contributions including:

- A industrial multimodal architecture framework (in Figure 1) for scalable ingestion, captioning, and retrieval of large amount of data.
- Four key findings from industrial deployment: (1) multimodality boosts VLM performance across most tasks (8/12); (2) attention mechanisms show GPU-specific trade-offs; (3) both open- and closed-source models struggle with temporal alignment and keyframe reasoning; and (4) long-video splitting & compression is more error-prone rather than efficient.
- Extension of Video-MME with new subtasks (summarization, keyframe evaluation) along with newly designed evaluation schema using the Knowledge Graph, shown in Alg. 1, 2, 3.

2 Related Work

Vision-Language Models (VLMs) have advanced significantly, with several state-of-the-art models demonstrating strong performance across multimodal tasks (Fu et al., 2024a). Qwen-VL (Bai et al., 2023), improved vision–language alignment through extensive pretraining and data scaling. Recent multimodal Open-source models, such as CLIP (Radford et al., 2021), BLIP (Li et al., 2022) , FLAVA (Singh et al., 2022), and OFA (Wang et al., 2022), have demonstrated strong zero-shot and fine-tuned performance in various vision-language tasks across recognition, captioning, and retrieval. More recent VLMs—including Gemini Pro/Flash (Gemini Team, Google: Petko Georgiev and 1135 other authors, 2024), AdaReTaKe (Lourentzou et al., 2021), and Qwen2-VL (Bai et al., 2023), have pushed the boundaries of video-based multi-modal understanding by improving reasoning, temporal alignment, and multi-modal fusion strategies. As shown in recent leader-boards (Fu et al., 2024a), Qwen2-VL (Bai et al., 2023) achieves state-of-the-art results in long & short video caption tasks, reinforcing its role as a leading open-source VLMs.

Meanwhile, several benchmarks have been proposed to systematically evaluate the capabilities of MLLMs. Video-MME (Fu et al., 2024a) introduces an evaluation framework specifically designed for VLM, evaluating their understanding of dynamic and multimodal content. Additionally, MME (Fu et al., 2023) provides a comprehensive benchmark for the evaluation of more general multimodal LLMs, while MME-Survey (Fu et al., 2024b) offers a detailed review of existing evaluation methodologies. The MME-RealWorld benchmark (Zhang et al., 2024a) further extends this evaluation to real-world, high-resolution scenarios, testing the robustness and generalization of multimodality beyond synthetic datasets.

What's more, efficient attention mechanisms have been crucial for scaling large multimodal models (Face, 2024). FlashAttention (Dao, 2023) and Scaled Dot-Product Attention (SDPA) (Vaswani et al., 2017) have played significant roles in improving efficiency in transformer-based architectures. FlashAttention reduces memory overhead and computational costs by optimizing key-query-value matrix operations, making it well-suited for large-scale applications. Similarly, SDPA, widely implemented in frameworks like Hugging Face's Transformers, optimizes inference performance on GPUs, particularly with hardware accelerations e.g. AMD ROCm and NVIDIA TensorRT (Face, 2024).

3 Dataset And Experimental Settings

Table 1: Statistics of Our Property Dataset.

Category	Details
VLM Models Covered	42, GPT series include GPT-4, Gemini 1.5 Pro. 2.0-Flash, Qwen-7B-VL, Qwen-72B, Owen VL Max, LLaVA-Video, Oryx-1.5, InternVL 2.5, Aria VideoLLaMA series, VideoChat Flash, NViLA, GPT-4o, Claude 3.5 Sonnet, TimeMarker, MiniCPM-V 3.2, MiniCPM-V 2.6, InternVL series, ST-LLaMA Video-XL, VITA-1.5, Kangaroo, Video-CCAM, ShareGemini, SIMM, Chat-Uni-VL 1.5, VideoChat2 Mistral, ShareGPT-4VVideo,
ALM Models Covered	Whisper-turbo and Whisper-large V2
Number of Videos	Over **25,326**.
Number of Audios	Over **888**.
Covered Variants	Over **14** Diseases Ares. From Nephrology, Ophthalmology, Oncology, ... to Hematology, Immunology, Dermatology.
Covered video format Types	**8**. MP4, M4V, QuickTime, WMV, WebM, MSVideo, MPG, and 3GPP
Covered audio format Types	**4**. '.mp3', '.wav', '.m4a', '.flac'
Covered Video Lengths	< 2 mins to over 3 hours
Language Types	Over 20 languages, including German, Italian, English, Mandarin Hokkien, Hindi, Korean, French, Dutch, Spanish, and more.

Here, we primarily adopt two well-established MLLM benchmarks, Video-MME (Fu et al., 2024a) and MMBench (Liu et al., 2023), along with evaluations of more than 40+ VLMs, as well as our pharmacy property dataset shown in Table 1, and Figure 4. The benchmarks allow standardized comparisons, while our dataset provides a realistic testbed for long-form industrial content.

Video-MME (Fu et al., 2024a) is the first full-spectrum multi-modal evaluation benchmark designed specifically for video-based MLLMs. It stands out from existing benchmarks with several key features: (1) Diversity in video types, covering six primary visual domains with 30 subfields to ensure broad scenario generalizability; (2) Temporal coverage, including short-, medium-, and long-term videos ranging from 11 seconds to 1 hour. Video-MME includes 900 manually selected videos totaling 254 hours, annotated with 2,700 question-answer pairs.

MMBench (Liu et al., 2023) is designed to assess our findings across diverse visual understanding tasks, including video recognition, captioning, visual question answering (VQA), and reasoning. Furthermore, our findings are tested from our industry data, includes 25,326 videos across eight formats (MP4, M4V, QuickTime, WMV, WebM, MSVideo, MPG, 3GPP) and 888 audio files across four formats (.mp3, .wav, .m4a, .flac). The content spans 14 disease areas, including oncology, hematology, immunology, ophthalmology, neuroscience, dermatology, nephrology, and respiratory disease, covering >20 languages. Video lengths range from under 2 minutes to more than 3 hours, reflecting the diversity of clinical trial recordings, medical lectures, and patient education materials.

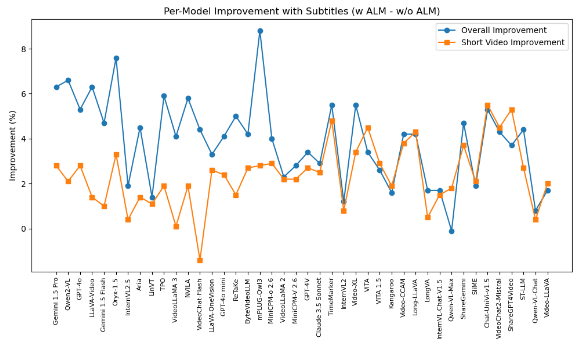
We benchmarked 42 VLMs / ALMs, including GPT-4 series (Ouyang et al., 2022), Gemini 1.5 Pro (Gemini Team, Google: Petko Georgiev and 1135 other authors, 2024), InternVL-Chat-V1.5 (Chen et al., 2024), and LLaVA-NeXT-Video (Zhang et al., 2024b), and Whisper etc. Each model was tested under default configurations and resource constrained GPU environments (NVIDIA A100 vs. A10G). Prompts are listed in Appendix Table 11.

4 Business And Technical Impact

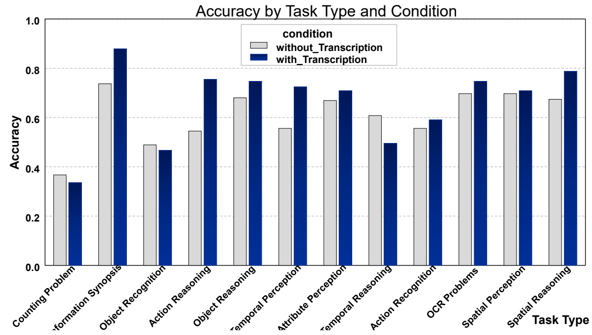
Finding reliable content remains a major challenge for healthcare professionals (HCPs) and patients. Traditional search methods are inefficient, leading to under-utilized assets and duplicated content creation. VLM-based NL search system streamlines discovery, reuse, and accessibility of video content. In production pilots, the system reduced the time required to create patient-facing materials by ***66%***, accelerating workflows and improving consistency. On Video-MME, the end-to-end processing time, from voice abstraction by Ffmpeg, Whisper Turbo transcription to Bedrock LLM improvement on VLM video summary & captions storage in the database, averaged 2.2 minutes per longer video, 1.7 minutes per medium-length video, 1.6 minutes per short video. Compared to manual inspection (252.5 hours for the dataset), this represents a ***94.4%*** reduction in effort for long videos and a ***88%*** reduction for overall video categories. Scaling to the entire dataset implies a savings of approximately c.a. ***224.3 hours***.

5 Main Results

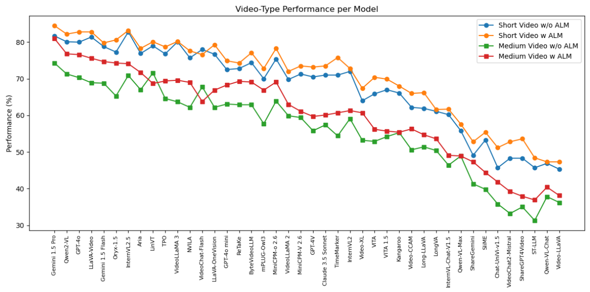
In this section, we first evaluate with the top two benchmarks, Video-MME (Fu et al., 2024a) and MMBench (Liu et al., 2023), as well as our proprietary dataset for multimodal vision language



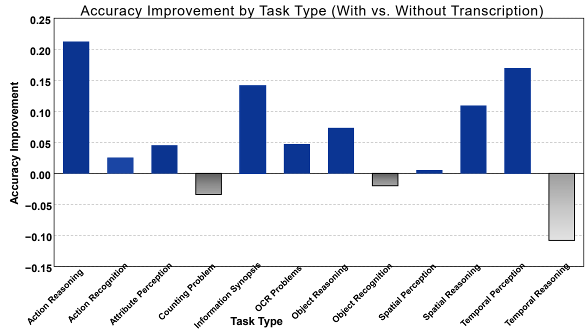
(a) Per-model improvement with ALM (overall vs. short).



(b) Accuracy by Subtask Type with/without Transcription.



(c) Video-type performance per model.



(d) Accuracy by Subtask Type with/without Transcription.

Figure 2: Multimodality matters. Combining metadata and voiceovers using LLM/ALMs could improve VLM summaries and understanding. Especially effective for longer videos, and for tasks like action recognition, object reasoning, and OCR, but may negatively impact temporal reasoning tasks.

Task Type	With ALM	Without ALM	Δ
Action Reasoning	0.759	0.545	+0.213
Action Recognition	0.589	0.564	+0.025
Attribute Perception	0.671	0.646	+0.025
Counting Problem	0.337	0.372	–0.035
Information Synopsis	0.879	0.737	+0.142
OCR Problems	0.744	0.698	+0.046
Object Recognition	0.469	0.490	–0.021
Spatial Perception	0.708	0.704	+0.005
Spatial Reasoning	0.789	0.680	+0.108
Temporal Perception	0.733	0.563	+0.171
Temporal Reasoning	0.500	0.615	–0.115
Average	0.683	0.626	+0.057

Table 2: Effect of multimodality on VLM performance (Video-MME) and task-level impact of audio transcription. Adding ALM-based voice transcriptions improves 8/12 task domains. Largest gains: Action Reasoning (+0.213), Information Synopsis (+0.142), Temporal Perception (+0.171). Negative effects: Counting (–0.035), Object Recognition (–0.021), Temporal Reasoning (–0.115).

Model	w/o	w/	Δ
Gemini 2.5 Pro	84.7	85.2	+0.5
Gemini 1.5 Pro	75.0	81.3	+6.3
Qwen2-VL	71.2	77.8	+6.6
GPT-4o	69.0	77.2	+8.2
LLaVA-Video	76.0	76.9	+0.9
Gemini 1.5 Flash	72.6	75.0	+2.4
Oryx-1.5	67.3	74.9	+7.6
InternVL2.5	67.6	74.0	+6.4
Aria	70.3	72.1	+1.8
LinVT	65.6	71.7	+6.1
TPO	66.2	71.5	+5.3
Average	68.4	72.3	+3.9

Table 3: Model-level impact of incorporating audio transcriptions. Overall accuracy increases from 58.4% to 62.3% (+3.9). Gains are largest for GPT-4o (+8.2), Oryx-1.5 (+7.6), and Qwen2-VL (+6.6). Newer models such as Gemini 2.5 Pro show smaller gains (+0.5), reflecting diminishing returns as models advance.

models (MVLMs), focusing on widely recognized 40 + vision language models (VLMs). Results are structured around four interesting findings.

Multimodality Matters. Combining metadata and voiceovers using LLMs/ALMs enhances VLM summaries and understanding, although ALM-only pipelines can already achieve competitive content summarization. Across 42 evaluated VLMs, including GPT-4, we observe a consistent positive trend in performance when audio transcriptions are incorporated. Notably, this improvement strongly correlates with **video length**. Overall accuracy increased from 58.4% to 62.3% (+3.9 points). For short videos, accuracy rose from 67.7% to 70.0% (+2.3 points), for medium videos from 69.3% to 74.3% (+5.0 points), and for long videos from 61.7% to 69.6% (+7.9 points). As video length increases, video Q&A Reasoning accuracy declines significantly from over 80% on short clips to below 50% on long-form videos, highlighting the persistent challenge of long video understanding. Incorporating subtitles and audio information mitigates this issue, yielding improvements of up to 7.9% for long videos, as shown in Figure 2 (a,b,c).

When further examining multimodality with subtask improvements, we can see that it improves 8/12 task domains, such as action reasoning, action recognition, information synopsis, OCR problems, and object reasoning, with improvements ranging from 2% to 20%. Here, we see the benefits of voice information in understanding procedural or sequential tasks. However, **temporal-related tasks remain challenging**, with several models showing reduced performance when processing both audio and visual inputs simultaneously (Figure 2d). This suggests that current VLMs still struggle with synchronizing multimodal context over extended time spans. Cross-modal attention mechanisms are typically optimized for semantic alignment rather than temporal synchronization. Our results suggest that while multimodal inputs improve overall understanding, they can introduce temporal noise, leading to higher rates of misaligned timestamps and incorrect segment boundaries in long videos (see Appendix Table 6). Additionally, we also noticed a decline in completeness of answered questions, dropping from 817 to 699, as CUDA out-of-memory (OOM) issues frequently occurred when reasoning with both audio & visual information.

Attention Mechanisms Matter. The VLM community has shown growing interest in FlashAttention-2 (Dao, 2023), as it provides an efficient implementation of the standard attention mechanism by parallelizing computation across sequence length and optimizing GPU memory usage. While FlashAttention significantly accelerates inference, most leading VLMs are trained and benchmarked on high-end GPUs such as the NVIDIA A100, which inherently benefits from optimized tensor operations and high memory bandwidth.

However, **FlashAttention-2 also exhibits strict GPU architecture dependencies**, requiring `Torch` $>$ 3.6 and being optimized primarily for NVIDIA A100 or select AMD GPUs (e.g., Instinct MI210, MI250). In addition, dtype constraints limit its operation to `fp16` or `bf16`, which restricts deployment in many industrial environments. **Deploying these mechanisms on standardized A10G Tensor Core GPUs—rather than on large A100 UltraCluster setups—introduces major cost and compatibility challenges.** For instance, an A100 UltraCluster costs approximately $40.96 per hour, whereas AWS G5 instances (A10G GPUs) range from $5.67 to $8.14 per hour, making A100-based inference roughly eight times more expensive. Empirically, FlashAttention-2 performs best on A100 GPUs and remains a popular choice for research benchmarks. In contrast, Scaled Dot-Product Attention (SDPA) (Vaswani et al., 2017) proves to be a more practical option for A10G-based deployments, offering up to a 4× speed improvement. Notably, SDPA achieved higher accuracy (58.73%) than FlashAttention (54.81%) when evaluated at FPS = 0.1 on the 2,700 Q&A Video-MME benchmark. SDPA completed inference in 4 hours and 37 minutes, compared to 7 hours and 40 minutes for FlashAttention-2, as summarized in Table 4, 8, 9. These findings suggest that **VLMs exhibit strong GPU architecture and FPS dependencies**, implying that the optimal attention mechanism varies with the hardware environment. SDPA (Vaswani et al., 2017) can also invoke FlashAttention and other memory-efficient attention kernels when needed, and native SDPAsupport is now expanding in the `Transformers` library. This underscores that selecting a well-matched–rather than the newest–attention mechanism can yield superior efficiency under realistic industrial GPU constraints.

Time Localization Remains Challenging: We evaluated VLM performance across three dimensions: speed, cost, and output completeness. Our findings indicate that Gemini 2.5 Flash is slower and more expensive, requiring 7–8 hours to process 300 videos at a cost of Euro 46.73 on Google Cloud. In contrast, Qwen-7B (Bai et al., 2023) demonstrated substantially faster processing, completing each video in 1 mins on an AWS EC2 `g5.24xlarge` instance, priced at $8.14 per hour. However, when running at 0.01 FPS with SDPA attention, Qwen-7B required approximately 3–4 hours to process 300 videos, as shown in Figure 3. In terms of **output completeness**, Gemini 2.5 Flash outperformed Qwen-7B, generating valid summaries for 297/300 videos (with acc.94.6%), while Qwen-7B produced only 74 valid summaries (with acc.74.3%). Additionally, **key frame extraction remained challenging for both models**, with Gemini 2.5 Flash achieving a keyframe accuracy of 35.1% (26/74), while Qwen-7B reached only 5.4% (4/74). For long videos, Gemini 2.5 Flash required roughly 3 mins per video, while Qwen-7B remained below one minute

Table 4: Comparison of speed, accuracy, and completeness across leading attention mechanism on all length videos.SDPA yields higher accuracy and 4× faster runtime on commodity GPUs (A10G), while FlashAttention favors high-end A100 GPUs.

Experiments	Processing Time	Total Answered (%)	Correct Answered (%)
SDPA (0.1 FPS)	4h 37m 2s	37%	58.73%
SDPA (0.01 FPS)	2h 12m 37s	87%	48.40%
FlashAttention (0.1 FPS)	7h 40m 17s	100%	54.81%
FlashAttention (0.01 FPS)	1h 50m 11s	100%	48.55%

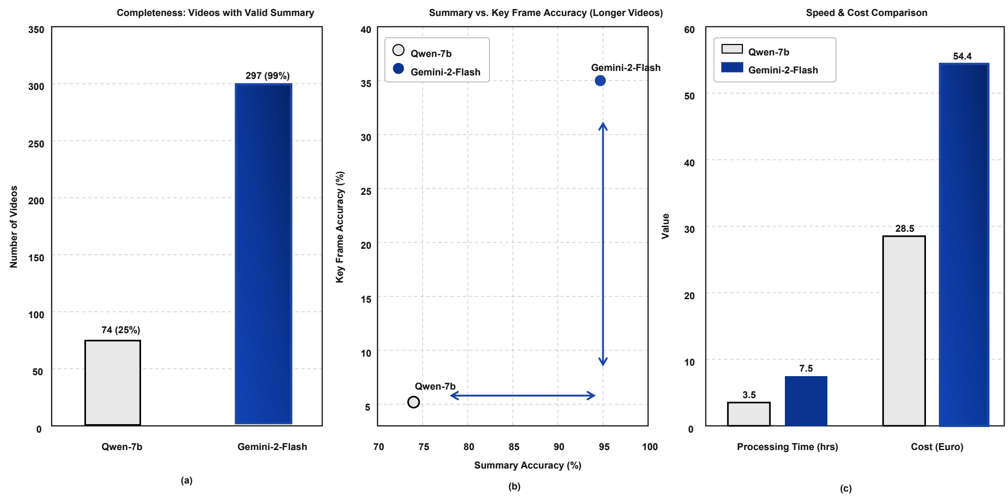


Figure 3: Time Localization Challenges for Open-Source and Closed-Source VLMs. Both top open-source and commercial models struggle with key frame detection, showing low accuracy (5-35%) and incorrect timestamps. Summaries are much more accurate, ranging from 75-95%.

per video. Despite these differences, **both leading open- and closed-source VLMs struggle with accurate temporal reasoning**, particularly in maintaining keyframe alignment and coherence across long video sequences (see Table 6). This highlights a persistent limitation in time-dependent video understanding for current VLM architectures.

Trade-off of Long Video Splitting. Processing long videos poses significant challenges, particularly on GPU-constrained instances. Out-of-memory (OOM) errors occur frequently, even when reducing frames per second (FPS) or lowering video resolution. A common mitigation strategy involves video splitting, compression, cutting. However, our analysis shows that these methods do not yield meaningful speed improvements and instead introduce additional challenges related to temporal alignment. **Splitting videos into multiple segments disrupts temporal consistency, making it increasingly difficult for VLMs to maintain coherent event sequences.** Moreover, compressing video files using standard tools such as FFmpeg requires substantial preprocessing time. Although shorter segments can be processed individually, the lack of contextual continuity causes VLMs to **focus on superficial cues**—such as logos, text color, or general stylistic attributes—rather than on semantic content. This leads to redundant descriptions and weaker keyframe alignment, as shown in Table 7.

6 Conclusion

In this work, we introduce an industrial framework for large-scale VLM-based video processing and NL search. Unlike prior studies proposing new architectures, our contribution lies in benchmarking, scaling, and analyzing existing VLMs under realistic GPU and compliance constraints. Our evaluation yields four key findings: (1) **Multimodality matters**—incorporating ALMs, transcriptions, and metadata improves video reasoning beyond static or bi-modal setups; (2) **Attention mechanisms matter**—matching attention to GPU architecture improves efficiency; (3) **Open- and closed-source VLMs** perform well in video summarization but still struggle with keyframe detection and timestamping; and (4) **Splitting long videos** often increases runtime and misalignment errors rather than improving efficiency. Beyond these insights, we extend Video-MME with new subtasks (summarization, keyframe extraction) and propose a knowledge-graph evaluation schema. Future research should exploring financial and manufacturing video domains under constrained hardware.

7 Limitations

We want to emphasize that this work is not intended to introduce a new VLM architecture, but rather to empirically characterize the capabilities, limitations, and failure modes of existing models under realistic long-video and hardware constraints. A main focus of this study is the justification of the pharma specific benefits, e.g. how to scale VLMs usage in industry GPU constrained hardware setting for business application. We provide a baseline comparison of more than 42 VLMs using both the Video-MME benchmark and our proprietary dataset. Future work should extend this line of research to other regulated domains, such as financial services and manufacturing videos e.g. Figure 7, to further validate the generalization of our solution blueprint.

8 Acknowledgments

We sincerely thank Samik Adhikary and Puneet Srivastava for their sponsorship support from Roche. We also appreciate the insightful discussions and technical assistance provided by Janina Kummerfeldt, Philipp Müller, and Kathrin Schwan from Accenture, Jennifer McGuire's business support from Roche, as well as Aditya Kumar's paper experiment support from Involead. This platform, RICI, would not have been possible without their contributions. We further extend our gratitude to the backend engineering teams supported RICI development, as well as to the healthcare professionals (HCPs), testers, and Roche Lab users whose consistent feedback brought our Content Finder use cases to life and enabled continuous improvement. This paper serves as our academic contribution, where we formalized experiments and evaluation methodologies using academic benchmarks. Through this work, we aim to share industry lessons learned and valuable large-scale GenAI experiments in the pharmaceutical domain.

References

Ai central (american college of radiology) – fda-cleared imaging ai directory. https://aicentral.acrdsi.org/. Over 200 FDA-cleared imaging AI products cataloged; Accessed 2025-10-07.

Artificial intelligence (ai/ml)-enabled medical devices. https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-enabled-medical-devices. Official FDA list of AI-enabled medical devices; Accessed 2025-10-07.

Isic: International skin imaging collaboration archive. https://www.isic-archive.com/. Dermoscopic/clinical *image* repository widely used for AI benchmarking.

2018. Idx-dr: De novo summary (den180001). https://www.accessdata.fda.gov/cdrh_docs/reviews/DEN180001.pdf. U.S. Food and Drug Administration De Novo classification for autonomous DR detection.

Sharib Ali, Debesh Jha, Noha Ghatwary, and et al. 2023. A multi-centre polyp detection and segmentation dataset for generalisability assessment (polypgen). *Scientific Data*.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*.

K. Behara and et al. 2024. Ai in dermatology: a comprehensive review into skin disease identification.

Hanna Borgli and et al. 2020. Hyperkvasir, a comprehensive multi-class image and video dataset for gastrointestinal endoscopy. *Scientific Data*.

Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, Ji Ma, Jiaqi Wang, Xiaoyi Dong, Hang Yan, Hewei Guo, Conghui He, Botian Shi, Zhenjiang Jin, Chao Xu, Bin Wang, Xingjian Wei, Wei Li, Wenjian Zhang, Bo Zhang, Pinlong Cai, Licheng Wen, Xiangchao Yan, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. 2024. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *arXiv preprint arXiv:2404.16821*.

Tri Dao. 2023. Flashattention-2: Faster attention with better parallelism and work partitioning. *arXiv preprint arXiv:2307.08691*.

L. Dick and et al. 2024. Automated analysis of operative video in surgical training: a scoping review.

Dawei Du, Yuankai Qi, Hongyang Yu, Yifan Yang, Kaiwen Duan, Guorong Li, Weigang Zhang, Qingming Huang, and Qi Tian. 2018. The unmanned aerial vehicle benchmark: Object detection and tracking. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 370–386.

Hugging Face. 2024. Transformers performance inference on gpu. Accessed: 2024-03-10.

Ghazala Fatima, Imran Khan, Saeed Farooq, et al. 2024. A comprehensive review of advances in digital pathology. *Diagnostics*.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, et al. 2023. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*.

Chaoyou Fu, Yuhan Dai, Yondong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. 2024a. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. *arXiv preprint arXiv:2405.21075*.

Chaoyou Fu, Yi-Fan Zhang, Shukang Yin, Bo Li, Xinyu Fang, Sirui Zhao, Haodong Duan, Xing Sun, Ziwei Liu, Liang Wang, et al. 2024b. Mme-survey: A comprehensive survey on evaluation of multimodal llms. *arXiv preprint arXiv:2411.15296*.

Gemini Team, Google: Petko Georgiev and 1135 other authors. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

N. Ghamsarian and et al. 2024. Cataract-1k: A large-scale cataract surgery video dataset for surgical workflow analysis. *Scientific Data*.

Md. Haque et al. 2025. Demographic bias in public remote photoplethysmography datasets. *npj Digital Medicine*.

Bin Huang et al. 2023. Challenges and prospects of visual contactless physiological monitoring. *npj Digital Medicine*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2021. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Preprint*, arXiv:2005.11401.

Johann Li, Guangming Zhu, and Cong Hua. 2021. A systematic collection of medical image datasets for deep learning. *arXiv preprint arXiv:2106.12864*.

Junnan Li, Ramprasaath R. Selvaraju, Akhilesh D. Gotmare, Shafiq Joty, Caiming Xiong, and Steven C.H. Hoi. 2022. Blip: Bootstrapped language-image pretraining for unified vision-language understanding and generation. In *Proceedings of the 39th International Conference on Machine Learning (ICML)*, pages 12888–12900.

Qiang Li, Mingkun Tan, Xun Zhao, Dan Zhang, Daoan Zhang, Shengzhao Lei, Anderson S. Chu, Lujun Li, and Porawit Kamnoedboon. 2025a. How LLMs react to industrial spatio-temporal data? assessing hallucination with a novel traffic incident benchmark dataset. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 3: Industry Track)*, pages 36–53. Association for Computational Linguistics.

Zongxia Li, Xiyang Wu, Hongyang Du, Huy Nghiem, and Guangyao Shi. 2025b. Benchmark evaluations, applications, and challenges of large vision language models: A survey. *Preprint*, arXiv:2501.02189.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. 2023. Mmbench: Is your multi-modal model an all-around player? *arXiv:2307.06281*.

Ismini Lourentzou, Chen Chen, and ChengXiang Zhai. 2021. Adarenet: Adaptive reweighted semi-supervised active learning. In *Proceedings of the 14th ACM International Conference on PErvasive Technologies Related to Assistive Environments (PETRA)*, pages 1–8.

O. Mahmoud and et al. 2024. Catstep: Automated cataract surgical phase classification using deep learning. *Ophthalmology Science*.

Suyash Mishra, Qiang Li, Srikanth Patil, and Anubhav Girdhar. 2026. From understanding to engagement: Personalized pharmacy video clips via vision language models (vlms). *Preprint*, arXiv:2601.05059.

R. et al. Najjar. 2023. Redefining radiology: A review of artificial intelligence. *PMC*.

OpenAI. 2023. Chatgpt for intelligence briefing: A case study. Accessed: March 2025.

OpenAI. 2024. Chatgpt: Optimizing language models for dialogue. Accessed: March 2025.

T. Oshika and et al. 2024. Artificial intelligence applications in ophthalmology.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744.

Evrim Ozmermer and Qiang Li. 2023. Self-supervised learning with temporary exact solutions: Linear projection. In *2023 IEEE 21st International Conference on Industrial Informatics (INDIN)*, pages 1–7.

L. et al. Pinto-Coelho. 2023. How artificial intelligence is shaping medical imaging. *PMC*.

Xiaoye Qu, Jiashuo Sun, Wei Wei, Daizong Liu, Jianfeng Dong, and Yu Cheng. 2025. Look, compare, decide: Alleviating hallucination in large vision-language models via multi-view multi-path reasoning. In *Proceedings of the 31st International Conference on Computational Linguistics*. Association for Computational Linguistics.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, pages 8748–8763.

Manuel Sebastián Ríos and et al. 2023. Cholec80-cvs: An open dataset with an evaluation of strasberg's critical view of safety for ai. *Scientific Data*.

H. Saeidi and et al. 2022. Autonomous robotic laparoscopic surgery for intestinal anastomosis. *Science Robotics*.

Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Dhruv Mahajan, Xinlei Wu, Christoph Feichtenhofer, Trevor Darrell, and Ross Girshick. 2022. Flava: A foundational vision and language model for generalized multimodal learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15638–15650.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. 2022. Large language models encode clinical knowledge. *arXiv preprint arXiv:2212.13138*.

K. et al. Strzępek. 2023. Quantitative and qualitative analysis of agricultural fields with uavs. *PMC*.

Amanda Tan-Garcia et al. 2025. Computational pathology in the age of artificial intelligence. *Histopathology*.

Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, et al. 2022. Lamda: Language models for dialog applications. *arXiv preprint arXiv:2201.08239*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*.

D. et al. Wang. 2025. A survey of unmanned aerial vehicles and deep learning in precision agriculture. *ScienceDirect*.

Peng Wang, An Yang, Rui Men, Ming Zhou, Yinan Zhang, Junyang Lin, Xu Sun, Shuai Li, Houqiang Wang, and Luo Si. 2022. Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. *arXiv preprint arXiv:2202.03052*.

Jing Wei, Zhanqing Li, Alexei Lyapustin, Jun Wang, Oleg Dubovik, Joel Schwartz, Lin Sun, Chi Li, Song Liu, Tong Zhu, et al. 2023. First close insight into global daily gapless 1 km $pm_{2.5}$ pollution, variability, and health impact. *Nature Communications*, 14:8349.

X. Xu and et al. 2024. The application of artificial intelligence in diabetic retinopathy: progress and challenges.

Kaixun Yang, Mladen Raković, Zhiping Liang, Lixiang Yan, Zijie Zeng, Yizhou Fan, Dragan Gašević, and Guanliang Chen. 2024. Modifying ai, enhancing essays: How active engagement with generative ai boosts writing quality. *Preprint*, arXiv:2412.07200.

Moran Yanuka, Assaf Ben-Kish, Yonatan Bitton, Idan Szpektor, and Raja Giryes. 2025. Bridging the visual gap: Fine-tuning multimodal models with knowledge-adapted captions. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 10497–10518. Association for Computational Linguistics.

A. Yilmaz and et al. 2024. Derm12345: A large, multi-source dermatoscopic skin lesion dataset. *Scientific Data*.

Dan Zhang, Fangfang Zhou, Yuwen Jiang, Felix Albu, Yuanzhou Wei, Xiao Yang, and Qiang Li. 2026. Unleashing the power of self-supervised image denoising: A comprehensive review. *Computer Vision and Image Understanding*, 269:104786.

Dan Zhang, Fangfang Zhou, Yuwen Jiang, and Zhengming Fu. 2023. Mm-bsn: Self-supervised image denoising for real-world with multi-mask based on blind-spot network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 4189–4198.

Yi-Fan Zhang, Huanyu Zhang, Haochen Tian, Chaoyou Fu, Shuangqing Zhang, Junfei Wu, Feng Li, Kun Wang, Qingsong Wen, Zhang Zhang, et al. 2024a. Mme-realworld: Could your multimodal llm challenge high-resolution real-world scenarios that are difficult for humans? *arXiv preprint arXiv:2408.13257*.

Yifu Zhang, Peize Sun, Yi Jiang, Dongdong Yu, Fucheng Weng, Zehuan Yuan, Ping Luo, Wenyu Liu, and Xinggang Wang. 2022. Bytetrack: Multi-object tracking by associating every detection box. In *Computer Vision – ECCV 2022*, volume 13682 of *Lecture Notes in Computer Science*, pages 1–21. Springer, Cham.

Yuanhan Zhang, Bo Li, Haotian Liu, Yong Jae Lee, Liangke Gui, Di Fu, Jiashi Feng, Ziwei Liu, and Chunyuan Li. 2024b. Llava-next: A strong zero-shot video understanding model.

A Appendix

In this section we provide the supplementary compiled together with the main paper includes:

- Evaluation Metrics and Knowledge Graph Evaluation Schema on Algorithm 1, 2, 3, and Figure 5;
- Ablation study on Frame Per Second (FPS) in Table 8, Table 9;
- Property Dataset distribution on Table 10, and VideoMME raw rata example on Table 5;
- Deliverable attributes of each VLM / ALM and Metadata on Figure 6;
- The training details and hyper-parameters of experiments including prompts lists in Table 11, output example on Table 6, 7;
- The business value case and area of impact of GenAI-driven Video Processing on Table 12, Table 13 and Figure 7.

A.1 Evaluation Metrics

Assigned accuracy scores strategies in Finding 3, Time Localization Challenges in Open-Source and Closed-Source VLMs.

$$\text{Scores}_{a,g} = \frac{1}{n_a}\sum_{i=1}^{n_a} S(x) \tag{1}$$

where

$$S(x) = \begin{cases} 1 & \text{if } S_{a,i} = g_{a,i} \\ \\ 0 & \text{if } S_{a,i}! = g_{a,i} \end{cases}$$

where S is the Matching Node score, $a \in A$ refers to an Key Frame or Summary scenarios, g refers to ground truth of timestamp, and n_a is the total number of valid video output (e.g., if 74 videos have valid JSON outputs, we match key frames to verify timestamp accuracy and compare summary accordingly).

A.2 Summary & Key Frame evaluation using Knowledge Graph

To compare the quality of video summaries generated by VLMs, we then employ a knowledge graph-based method. This is particularly useful in scenarios where human-annotated ground truth is incomplete or unavailable, such as with large-scale video datasets in industry setting.

A.2.1 Knowledge Graph Construction

We use the `NetworkX` library with `DiGraph` to construct the knowledge graph, `NetworkX` library encapsulated so well where:

- Nodes represent extracted keyframes and conceptual entities (nouns or keywords) from the generated summaries.
- Edges represent semantic or temporal relationships between these concepts.
- The graph layout is generated using the `spring_layout` function, which implements the Fruchterman-Reingold force-directed algorithm, as shown in Algorithm 1.

A.2.2 Mathematical Basis Behind

The force-directed layout models the graph using physical analogies:

- **Repulsion**: All nodes repel each other according to **Coulomb's law**.
- **Attraction**: Connected nodes attract each other like springs **(Hooke's law)**.

These forces iteratively adjust node positions until a stable configuration is reached, visually revealing clustering and coverage. Thereby:

- **Node Count**: Indicates the richness or breadth of extracted information.
- **Node to Node Distance**: Measures how widely concepts or key words are spread in the graph.
- **Distance to Central Node**: We compute shortest path lengths using Dijkstra's algorithm to measure how far keyframe nodes are from the central summary node.as shown in Algorithm 2.

This is an emerging area with ongoing efforts to define metrics for summary evaluation without ground truth. Recent work from researchers at Google and Apple (Yanuka et al., 2025) (e.g., ***Descriptiveness Recall, Contradiction Precision***, Cosine Similarity) highlights the need for new metrics when ground truth of video summary, Key frame captions are missing.

In summary, we introduce new task domains based on open-source Video-MME (Fu et al., 2024a) tasks, which previously lacked summary and keyframe subtasks due to the manual effort required for key frame localization labeling. To address this, we propose a knowledge graph approach to compare the output quality of various VLMs, as shown in Algorithm 3. This approach provides visually interpretable and computationally supported method to assess summary & key frame quality using graph-based representations, forming the basis for future work on automated evaluation metrics in GenAI applications.

Algorithm 1 Force-Directed Graph Layout (***Fruchterman-Reingold)*** conceptually

Let:
    $d(u, v) \quad \leftarrow$ distance between nodes u, v
    $k \leftarrow C \cdot \sqrt{\frac{A}{n}}$ ▷
    C is constant, A is layout area, n is number of nodes, K is optimal distance between nodes.
Forces:
    1. Attractive force (between connected nodes):

$$F_{\text{attr}}(d) = \frac{d^2}{k}$$

    2. Repulsive force (between all nodes):

$$F_{\text{rep}}(d) = \frac{k^2}{d}$$

Loop: Apply forces iteratively until convergence or maximum iterations reached.

Algorithm 2 ***Dijkstra's Algorithm*** for Shortest Paths

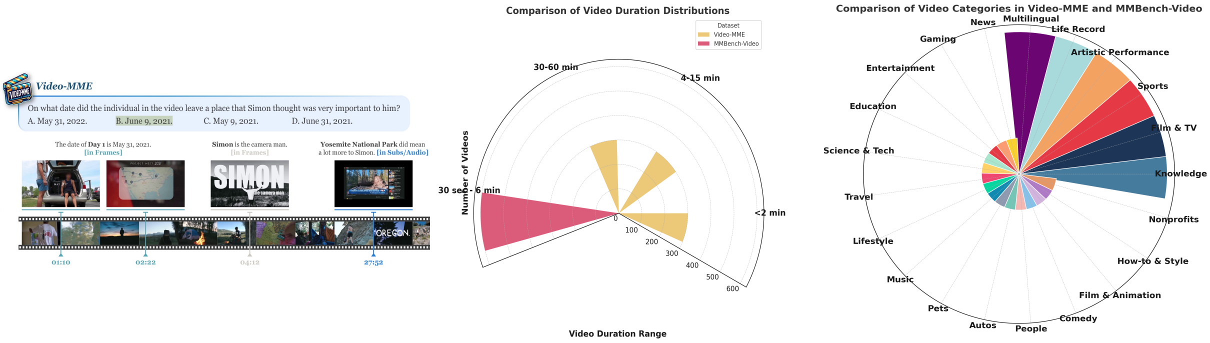
Input: Directed graph $G = (V, E)$ with non-negative weights $w(u, v) \geq 0$
Input: Source node s
Initialization:
    distance$[v] \leftarrow \infty$ for all $v \in V$
    distance$[s] \leftarrow 0$
    Initialize a priority queue Q
while Q is not empty **do**
    Extract node u with minimum distance$[u]$
    for each neighbor v of u **do**
        Update:

$$\text{dist}[v] \leftarrow \min(\text{dist}[v], \text{dist}[u] + w(u, v))$$

    end for
end while
Output: Shortest distances from s to all $v \in V$

Algorithm 3 Knowledge Graph Construction for ***Summary and Key Frame Evaluation***

Input: JSON data with key frames from Gemini-2 Flash and Qwen-7B
Output: Visualized Knowledge Graph
Step 1: Initialize Graph
Create directed graph $G \leftarrow$ nx.DiGraph(). Nodes (V) $\rightarrow$ Individual entities in the graph (e.g., "Gemini-2 Flash", "Snow White in rags"). Edges (E) $\rightarrow$ Directed connections between nodes (e.g., "Gemini-2 Flash" $\rightarrow$ "Key Frames"). Attributes $\rightarrow$ Additional properties of nodes/edges (e.g., color, size).
DiGraph G = (V, E) is defined as: V=nodes, E=(source, target), where each edge has a direction.
Step 2: Add Core Nodes
Add node $KeyFrames$ with attributes (color: gray, size: 800)
Add node $VideoSummary$ with attributes (color: gray, size: 600)
Step 3: Connect Models to Core Nodes
Add edge $(e.g., KeyFrames)$
Add edge $(e.g., VideoSummary)$
Step 4: Add Key Frames for Each Model
for each $(timestamp, description)$ in **r.g. Gemini-2 Flash key frames do**
    Add node $description$ with attributes (color: light blue, size: 400)
    Add edge $(KeyFrames, description)$
end for
Step 5: Visualize the Graph
Compute node layout $pos \leftarrow$ spring_layout(G, seed=42)
Extract node colors and sizes for Knowledge Graph: Key Frames Evaluation.



(a) (b) (c)

Figure 4: Comparison of the Video-MME (Fu et al., 2024a) and MMBench-Video datasets (Liu et al., 2023) in terms of video categories and duration distributions. The Video-MME dataset consists of 900 videos spanning six primary visual domains with 30 subfields, categorized into 300 short-term (<2 min), 300 medium-term (4-15 min), and 300 long-term (30-60 min) videos. In contrast, the MMBench-Video dataset comprises approximately 609 videos across 16 major categories, with durations ranging all from 30 seconds to 6 minutes.

Table 5: Video-MME (Fu et al., 2024a) raw dataset structure, Q/A details in Json format.

Field	**Details**
Video ID	001
Duration	Short
Domain	Knowledge
Sub-Category	Humanity & History
URL	`https://www.youtube.com/ watch?v=fFjy93ACGo8`
VideoID	fFjy93ACGo8
Question ID	001-2
Task Type	Information Synopsis
Question	What is the genre of this video?
Options	
A	It is a news report that introduces the history behind Christmas decorations.
B	It is a documentary on the evolution of Christmas holiday recipes.
C	It is a travel vlog exploring Christmas markets around the world.
D	It is a tutorial on DIY Christmas ornament crafting.
Answer	**A**

Table 6: Comparison of Video Summarization: Gemini-2-Flash vs Qwen-7B. Here *the timestamps are all wrong*.

Category	**Gemini-2-Flash**	**Qwen-7B**
Video ID	P69idA8JO98	P69idA8JO98
Duration	Long	Long
Domain	Artistic Performance	Artistic Performance
Summary	A stage performance of *Snow White*. The Evil Queen consults the Magic Mirror, instructs Snow White to clean the castle, and the story unfolds as Snow White meets the Seven Dwarfs, receives the poisoned apple, collapses, and is revived by the Prince.	A fairy tale performance, likely *Snow White and the Seven Dwarfs*. The video introduces characters, a forest scene, a confrontation between a queen and a prince, interactions between Snow White and the dwarfs, and ends with a song.

Key Frames (Gemini-2-Flash)

Time	**Description**
00:08	Magic Mirror reveals an angry face
00:42	Snow White in rags looking at her stepmother
00:46	The Evil Queen on a castle balcony
01:27	Dopey Dwarf dancing in silk costume
01:37	Ethereal dancer twirling with a deer
02:17	Snow White with basket approaching animals
02:57	Snow White collapses onto a stage of rocks
03:42	Snow White at a wishing well
05:09	The Evil Queen on a balcony speaking to a soldier
05:37	Snow White dancing in her new dress
06:09	Snow White and her prince hold hands
07:02	Snow White falls, animals mourn her
07:26	The Prince awakens Snow White with a kiss
08:00	Snow White is held up for celebration
08:07	Evil Queen standing on castle balcony
09:05	Snow White lies in a glass coffin as prince kneels

Key Frames (Qwen-7B)

Time	**Description**
00:00	Introduction of characters and setting
02:00	Scene with group of people in a forest
04:00	Confrontation between a queen and a prince
06:00	Introduction of the dwarfs as Snow White's friends
08:00	Scenes of the dwarfs working and interacting with Snow White
10:00	Snow White singing a song with the dwarfs

Table 7: With and Without ALM for Summary and Key Frame: Using ALM for voiceover transcription improves the relevance and precision of the summary content *(see the blue marked texts)*. However, timing remains a significant issue.

Component	Output Text
Prompt (Qwen Model)	"Focusing on the content and workflow rather than specific logos or the color of text? After summarizing, list the key frames with brief captions in the format (00:00, caption). Ensure the analysis is accurate and avoid including any assumptions or extrapolations. Use a pharmaceutical expert domain perspective to enhance relevance and precision."
Whisper Model (Voice-over Transcription)	" "id": 0, "start": 7.72, "end": 13.6, "text": " PCR of course refers to pathological complete response where once the patient has surgery" ... "**Full Transcription**": " PCR of course refers to pathological complete response where once the patient has surgery the pathologist does not find any cancer at all and pleasingly over the last sort of 15-20 years we've seen improvements in systemic treatment to such an extent that certainly for HER2 positive breast cancers we are now able to expect 50-60% of patients who have a PCR following the neoadjuvant treatment and we know that the patients who have a PCR are the ones who are likely to do the best in the long term. "
Qwen Model Output (Before Transcription)	**VLM Output**: "The video appears to be a medical demonstration, likely related to the Roche brand, which is known for its involvement in healthcare and diagnostics. The video showcases a 3D model of a human body with various organs highlighted, including the liver and brain. The model is displayed within a transparent box, possibly representing a medical device or a simulation environment. **Key Frames with Captions**: 00:00 - Introduction: The video begins with a close-up of a transparent box containing a 3D model of a human body. 01:00 - Organ Highlighting: The liver and brain are highlighted in the model, indicating a focus on these organs. 02:00 - Interactive Display: The model rotates, providing a better view of the highlighted organs. 03:00 - Data Overlay: Digital text and graphics appear, likely representing medical data or information related to the highlighted organs. 04:00 - Conclusion: The video ends with a summary or conclusion, possibly emphasizing the importance of the highlighted organs or the role of the device in medical diagnostics. The video likely aims to educate viewers about the importance of the liver and brain in the human body and how the device or simulation can be used to study or diagnose related conditions." "
Qwen Model Output (After Transcription Integration)	**VLM Output**: "Here is a summary focusing on the content and workflow from a pharmaceutical expert's perspective, along with key frames and brief captions" Summary: The video provides an explanation of the term "*pathological complete response" (pCR)* and its significance in cancer treatment, particularly for *HER2-positive breast cancer*. The speaker, a medical professional, *defines pCR as the absence of any detectable cancer cells in the surgical specimen after neoadjuvant (pre-operative) systemic treatment*. He highlights the remarkable progress made in systemic therapies over the past 15-20 years, *enabling 50-60% of HER2-positive breast cancer patients to achieve pCR following neoadjuvant treatment*. The speaker *emphasizes that patients who achieve pCR have the best long-term outcomes*, underscoring the importance of this endpoint in cancer management. **Key Frames**:(00:07) Introduction to the concept of pathological complete response (pCR).(00:19) Progress in systemic treatments improving pCR rates. (00:24) Specifically mentioning 50-60% pCR rates in HER2-positive breast cancer.(00:33) Linking pCR to better long-term outcomes for patients."

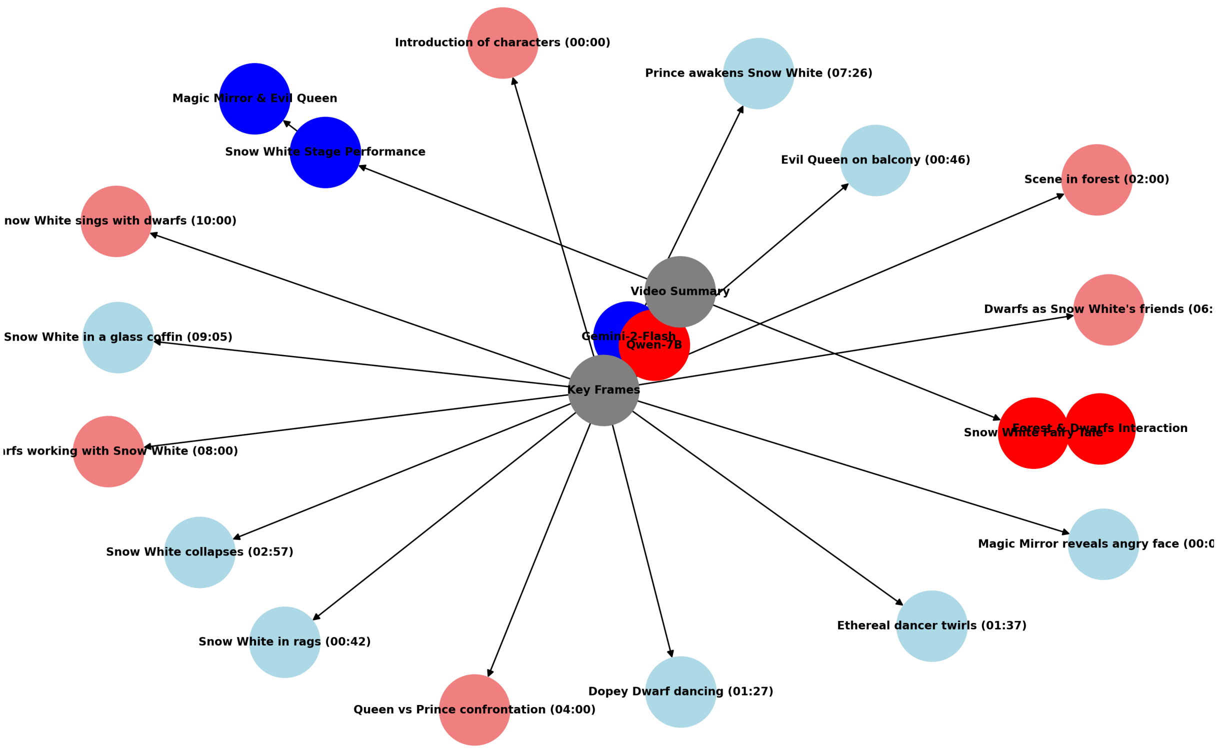


Figure 5: Knowledge graph for summary and key frames using Qwen and Gemini models. The knowledge graph visualizes the comparison between Gemini-2-Flash and Qwen-7B in summarizing a 'Snow White' stage performance. Each key frame from Gemini-2 Flash is marked in light blue and video summary in dark blue, while Qwen-7B's key frames are in light red, and video summary in red. The central node represents the key frames, with connections showing their relationships to each model's summary. Gemini-2-Flash emphasizes narrative elements such as the Magic Mirror, the Evil Queen, and the climax involving Snow White's revival, while Qwen-7B structures the story around broad thematic transitions like character introductions, forest scenes, and musical elements. This graph presents a structured comparison of the keyframes extracted by Gemini-2-Flash and Qwen-7B from a 'Snow White' performance. *The blue nodes represent Gemini-2-Flash's emphasis on theatrical storytelling, focusing on individual character moments, while the red nodes highlight Qwen-7B's broader narrative structure*, including interactions between Snow White and supporting characters. Additionally, the red nodes are more widely distributed, whereas the blue nodes are clustered more closely, indicating a difference in granularity and focus.

Table 8: Ablation Study on FPS (1): Attention Mechanism Dependence on FPS. Comparison of Speed, Accuracy, and Completeness Across Leading Attention Mechanisms on Short Videos (<120s). *This study highlights the strong dependence of FPS on each model's performance. For short videos, FlashAttention is recommended over SDPA*. *Default settings as per the Video-MME benchmark (FPS = 1), with no additional audio ALM transcription fed into the Qwen model.

Experiments on Short Videos*	**Processing Time**	**Total Answered (%)**	**Correct Answered (%)**
SDPA (1 FPS)	44m 12s	6%	64.81%
FlashAttention (1 FPS)	1h 30m 12s	100%	70.78%
Experiments on All Length Videos	**Processing Time**	**Total Answered (%)**	**Correct Answered (%)**
SDPA (0.1 FPS)	4h 37m 2s	37%	58.73%
FlashAttention (0.1 FPS)	7h 40m 17s	100%	54.81%
SDPA (0.01 FPS)	2h 12m 37s	87%	48.40%
FlashAttention (0.01 FPS)	1h 50m 11s	100%	48.55%

Table 9: Ablation Study on FPS (2) *Reducing FPS does not necessarily help the Qwen model answer more questions correctly*. In fact, it can have a negative impact, as lower frames per second lead to missing information. Here, the completeness percentage increases from 37% to 87% significantly, but the accuracy drop from 58.73% to 48.40%. However, *with the support of audio ALM transcription, accuracy is maintained*, improving from 58.73% to 61.80% when FPS = 0.1, and from 48.40% to 49.36%. This further validates our first finding from a different perspective.

Experiments on All Length Videos	**Total Answered (%)**	**Correct Answered (%)**
SDPA (0.1 FPS) without Audio Transcription	37%	58.73%
SDPA (0.1 FPS) with Audio Transcription	26.78%	61.80%
SDPA (0.01 FPS) without Audio Transcription	87%	48.40%
SDPA (0.01 FPS) with Audio Transcription	68.92%	49.36%

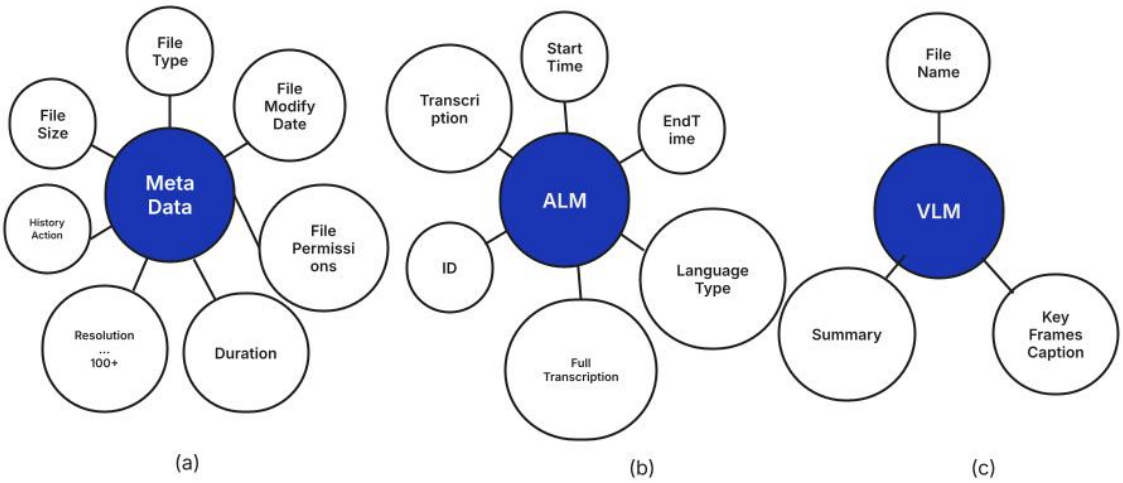


Figure 6: Deliverable Attributes. Define each GenAI output file attributes. *ALM stands for Audio Large Language model generated output, and *VLM represents Video Large Language model generated output.

Table 10: Distribution of Property Audio and Video Data Across Medical Diseases Specialties.

Specialty	**Audio**	**Video**	**Specialty**	**Audio**	**Video**
Oncology	208	8934	Ophthalmology	159	2862
Cardiovascular	1	14	Respiratory Disease	16	467
Dermatology	0	30	Nephrology	1	380
Hematology	67	3606	Not Applicable	59	2853
Immunology	144	510	Movement Disorder	9	289
Infectious Disease	1	239	Inflammatory Disease	20	222
Metabolism	0	6	Neuroscience	202	4914

Table 11: The backbones, hyper-parameters, and prompt settings of **selective** SOTA VLMs.

Model Description	Type	Token Limit	API Price in Dollars	Modality	Prompt Example
GPT-4 Turbo, The latest GPT-4 model with improved instruction, reproducible outputs, parallel function calling. Returns max of 4,096 output tokens. Training data up to Apr 2023	gpt-4-1106-preview	Input token limit:$128K$	Input $0.06/K$ Tokens. Output $0.12/K$ Tokens	Text Generation, Chat Completion, Image to Text	Could you please provide a summary of this video based on sample frames focusing on the content and workflow rather than specific logos or the color of text? After summarizing, list the key frames with brief captions in the format (00:00, caption). Ensure the analysis is accurate and avoid including any assumptions or extrapolations. Use an expert domain perspective to enhance relevance and precision. Do not repeat sentences or focus on QR codes or logos.
Qwen2-VL. updated on Huggingface Jan. 2025	Qwen2-VL-7B	Input token limit:$32K$	Opensource	Text Generation, Video to Text	Could you please provide a summary of this video, focusing on the content and workflow rather than specific logos or the color of text? After summarizing, list the key frames with brief captions in the format (00:00, caption). Ensure the analysis is accurate and avoid including any assumptions or extrapolations. Use an expert domain perspective to enhance relevance and precision. Do not repeat sentences or focus on QR codes or logos.
Gemini 2.0 Flash Model, released data 05th Feb 2025	Gemini 2.0 Flash 001	Input token limit:$1048K$ Output token limit: $8K$	Per 1M tokens in USD: API Cost 0.10 (text / image / video) 0.70 (audio). Output 0.40. e.g., 300 Long Videos from Video-MME costed 56.62	Audio, images, video, text, and PDF to text	Could you please provide a summary of this video, focusing on the content and workflow rather than specific logos or the color of text? After summarizing, list the key frames with brief captions in the format (00:00, caption). Ensure the analysis is accurate and avoid including any assumptions or extrapolations. Use an expert domain perspective to enhance relevance and precision. Do not repeat sentences or focus on QR codes or logos.

Table 12: Area of Impact 1: AI/VLM-driven Video Processing for Medical Diagnosis.

Category	Applications
Radiology and Imaging	• **MRI/CT Scan Video Processing**: Advanced AI can analyze full-length MRI or CT scans in motion (e.g., cardiac MRI or functional MRI), detecting anomalies faster than manual review. • **Ultrasound Interpretation**: AI-powered real-time video analysis can help with fetal health assessments, echocardiography, and liver disease detection.
Endoscopy and Surgery	• **Colonoscopy Polyp Detection**: AI can process hours of colonoscopy footage to detect polyps in real-time, improving colorectal cancer screening. • **Robotic Surgery Assistance**: AI-driven video processing can provide real-time insights to surgeons, flagging anomalies or suggesting procedural adjustments.
Neurology and Movement Disorders	• **Seizure and Tremor Analysis**: AI can analyze EEG-video recordings to classify epilepsy types. • **Parkinson's and ALS Monitoring**: AI can assess gait, facial expressions, and movement from patient videos for early diagnosis and tracking progression.
Ophthalmology and Dermatology	• **Retinal Scan Analysis**: AI models can process retinal scan videos to detect early diabetic retinopathy or macular degeneration. • **Skin Cancer Detection**: Dermatologists can use AI-enhanced dermoscopy video processing to detect melanoma more accurately.
Pathology and Microscopy	• AI can analyze continuous microscopy footage to identify cancerous cells, bacterial infections, or rare hematological conditions in blood samples faster than human pathologists.
Remote Patient Monitoring	• Wearable devices that record and process patient videos (e.g., heart rate monitors, movement trackers) can enable early diagnosis of conditions like heart arrhythmias or sleep apnea at lower costs.

Table 13: Area of Impact 2: AI/VLM-driven Video Processing for Drone/Satellite Data Analysis.

Category	Applications
Agriculture and Environment	• **Crop Health Monitoring**: AI-driven video analysis can quickly identify stressed crops, pest infestations, or nutrient deficiencies. • **Deforestation and Land Use**: Detecting illegal logging or monitoring ecosystem changes becomes faster and cheaper. • **Disaster Assessment**: Rapid damage assessment after hurricanes, earthquakes, or floods helps authorities respond effectively.
Infrastructure and Urban Planning	• **Road and Bridge Inspections**: AI can process high-resolution drone footage to detect cracks, erosion, or weak points. • **Traffic and Urban Planning**: Satellite video can track congestion patterns and optimize urban development.
Defense and Security	• **Surveillance and Threat Detection**: Automated analysis of drone/satellite feeds can detect anomalies, unauthorized activities, or suspicious movements. • **Border and Maritime Security**: Continuous video monitoring can identify smuggling, illegal crossings, or unauthorized vessel movements.
Healthcare and Epidemiology	• **Mosquito-Borne Disease Prevention**: Satellite video can help detect standing water bodies where mosquitoes breed, aiding in malaria/dengue prevention. • **Air Pollution and Public Health**: Fast video analysis can track pollution hotspots, correlating air quality data with disease outbreaks.

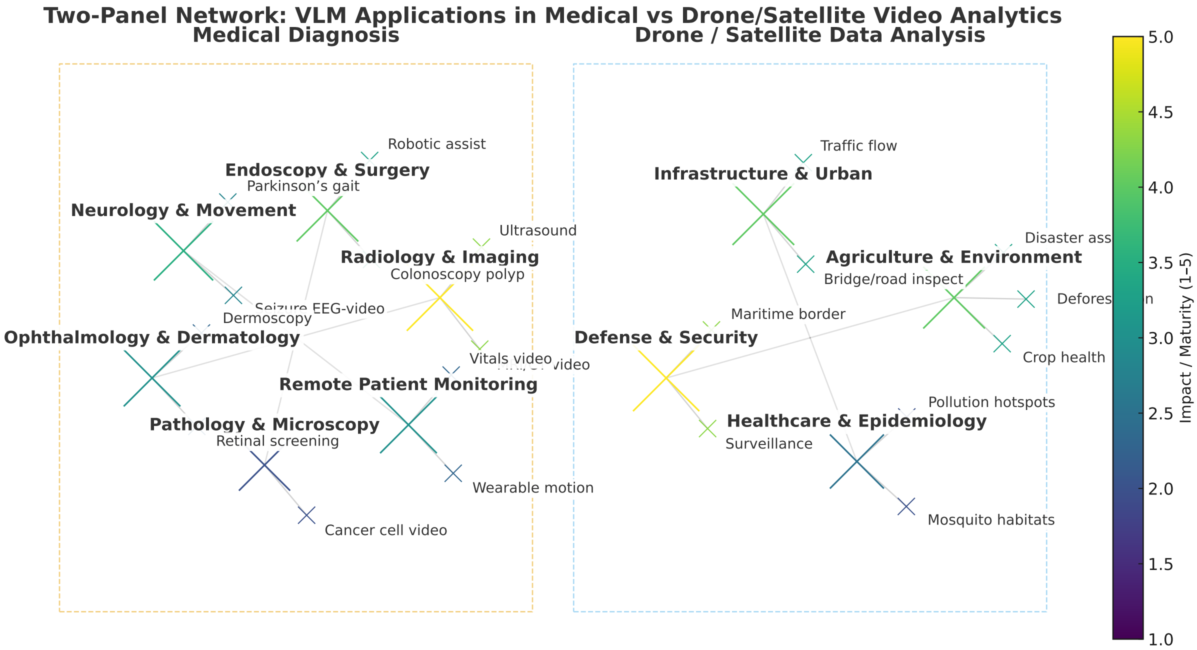


Figure 7: Two-Panel Network of Vision–Language Model (VLM) / AI Applications Across Medical and Drone/Satellite Video Analytics. Node color and size encode an *impact/maturity score* (1–5). Impact is estimated as a composite index: $\text{Impact} = 0.4R + 0.2D + 0.2A + 0.2C$, where R denotes normalized research intensity (publications indexed in PubMed, IEEE Xplore, Scopus from 2020–2025 containing "video + AI or VLM + domain"), D dataset availability (standardized/public video datasets), A application readiness (evidence of clinical or industrial deployment and/or regulatory signals), and C cross-domain generalizability (extent of transferability to other domains or modalities). Scores are normalized to $[1, 5]$ (5 = high maturity, well-validated, standardized datasets, active commercial use) and were assigned approximately as: Radiology & Imaging 5.0 with extensive literature (>200 papers, 2020–2025), benchmark datasets (e.g., MIMIC-CXR, cardiac MRI) (Li et al., 2021; Najjar, 2023; Pinto-Coelho, 2023), and multiple FDA-approved AI tools (FDA; ACR); Endoscopy & Surgery 4.0 with validated prototypes for polyp detection and robotic surgery support(Saeidi and et al., 2022), but moderate dataset availability (Borgli and et al., 2020; Ali et al., 2023; Ríos and et al., 2023); Neurology & Movement 3.5; Ophthalmology & Dermatology 3.0 with strong still-image AI base (fundus/OCT in ophthalmology; dermoscopy in dermatology), including FDA-cleared autonomous DR screening and large image datasets/challenges (FDA, 2018; Oshika and et al., 2024; Xu and et al., 2024; ISI; Yilmaz and et al., 2024; Behara and et al., 2024) but few real-time video pipelines available (Mahmoud and et al., 2024; Ghamsarian and et al., 2024; Dick and et al., 2024) ; Pathology & Microscopy 2.0 with only limited continuous video/time-lapse microscopy datasets and research mostly whole-slide image (WSI)–based (Fatima et al., 2024; Tan-Garcia et al., 2025); Remote Patient Monitoring 3.0 with active wearable and webcam-based monitoring studies and few open datasets (Haque et al., 2025; Huang et al., 2023) ; Agriculture & Environment 4.0 with large drone video datasets and real-world agricultural monitoring systems and high industrial uptake (Strzępek, 2023; Wang, 2025); Infrastructure & Urban 4.0 with drone/traffic inspection systems widely deployed and mature technical readiness (Du et al., 2018); Defense & Security 5.0 with advanced object tracking and anomaly detection systems in full industrial operation (Zhang et al., 2022) ; Healthcare & Epidemiology 2.5 with exploratory studies linking environmental video data with disease risk or pollution metrics (Wei et al., 2023). Smaller satellite nodes indicate representative sub-tasks within each category.