AIODT: Assisted Interactive Operator-in-the-loop Discovery Toolkit for Vietnamese Urban Video Retrieval

1st Given Name Surname
name of organization (of Aff.)
City, Country
email address or ORCID

2nd Given Name Surname
name of organization (of Aff.)
City, Country
email address or ORCID

3rd Given Name Surname
name of organization (of Aff.)
City, Country
email address or ORCID

4th Given Name Surname
name of organization (of Aff.)
City, Country
email address or ORCID

5th Given Name Surname
name of organization (of Aff.)
City, Country
email address or ORCID

***Abstract*—An operator asks a question about a large archive of city video in Vietnamese, the answer is one specific moment inside hundreds of hours of footage, and the binding constraint is time. Neither a fully automatic system nor a human operator meets that constraint alone: a Vietnamese query meets an English-trained dense backbone, no single channel carries the answer on its own, and self-hosted serving sits behind a hard network deadline. We present AIODT, a human-in-the-loop multimodal retrieval system hardened through the AI Challenge competition series. Weighted reciprocal-rank fusion combines five ranked lists, dense CLIP, caption embedding, and lexical search over captions, OCR, and ASR, a pointwise vision-language reranker re-judges the head of the list, and a Vietnamese-to-English bridge sits in front of the dense channel; the core runs on open weights on two data-center-class GPUs behind a public tunnel. The contribution is an *operational layer*: a console built for speed under pressure, with a one-hand keyboard loop, non-blocking evidence loading, neighbour-frame browsing, an assignment board for distributed annotation, and a submission builder. On the public half of each day's question set, counting only a day's best submission, we scored 9.8 of 13, 11.6 of 15, and 15.8 of 18, which is 75 to 88% of the attainable maximum. The interactive fast path returns under 1 s, a visual question-answering turn costs 3 s on the llama.cpp back-end and a 50-candidate pass 47 s on vLLM because prefill, not decode, dominates, and a 100 s ingress ceiling shapes every concurrency choice.**

***Index Terms*—multimodal AI for smart cities, AI applications in smart city services, large language models in urban applications, real-time intelligent systems, human-in-the-loop video retrieval**

I. Introduction

Cities generate video faster than people can watch it. Traffic-infrastructure cameras, port and warehouse cameras, delivery-fleet footage, urban news channels, and broadcast archives accumulate into archives of hundreds of hours, in which metadata cannot answer a typical operational question such as "find the moment the fire truck turns into the street with a Vietnamese signboard" or, from the organizer's sample set, "the launch of a private spacecraft, opening with four astronauts in black suits, one planned mission being the study of the polar aurora". The task is event-based retrieval: the system takes natural language as input and must return a specific frame at a specific timecode, from an archive too large to scan by hand and too diverse for a keyword index to catch.

Urban logistics and supply-chain systems need the same capability: monitoring cargo flow at ports and warehouses, tracing delivery fleets, and checking safety events on logistics video. Each task is an instance of "find one verbally described moment inside hundreds of hours". We take these domains as a *target deployment class* that the architecture generalizes to and *not* as measured data: we run the evaluation reported below on the AIC-2026 urban video corpus, which is broadcast news and public urban footage rather than private surveillance video, and every quantitative statement is bound to that scope.

Neither pole suffices in operation. A fully automatic system fails where a Vietnamese query meets an English-trained backbone, and where visual content leaves no textual trace. One person cannot scan hundreds of hours within the allotted time. Competitions such as the Video Browser Showdown [1], [2] have formalized this conclusion, since every top-ranked system there is a *human-machine* pair: the machine narrows the space and the human makes the final call. The engineering question is therefore not how to remove the human, but how to design the operational layer so that human and machine collaborate under time pressure.

Building such a pair for Vietnamese urban video raises four concrete challenges, and we met each as a measured constraint rather than an anticipated one. **C1, the cross-lingual gap.** The query arrives in Vietnamese while the strongest dense backbones are trained primarily on English, and the backbone we use accepts only 72 text tokens, so a language model must bridge and compress the long descriptive query without losing the discriminative term. **C2, no single channel is sufficient.** A visually described moment often leaves no textual trace,

whereas an on-screen caption or a spoken sentence sometimes carries the only evidence; we must therefore fuse the channels and re-judge the fused head, which costs time.

C3, serving cost under an infrastructure ceiling. Self-hosting the vision-language models is necessary for cost and reproducibility, but prefill rather than decode dominates a visual turn, so standard decode-side accelerations do not help. The deployment also sits behind a network layer that cuts any response past 100 s. **C4, operator throughput.** Once retrieval is good enough to place the answer somewhere in the list, the binding constraint moves to how many actions separate the operator from committing it, and to how several operators can share one question set without collision.

We present AIODT, a human-in-the-loop multimodal video retrieval system built and hardened through the Vietnamese AI Challenge series on the AIC-2026 corpus of 873 videos, and we address the four challenges in turn. A Vietnamese-to-English query bridge answers C1, and the console exposes that bridge for verification. Weighted rank fusion over five lists followed by a pointwise vision-language reranker over the head of the list answers C2. A two-back-end self-hosted serving composition answers C3, its concurrency tuned against the measured 100 s ceiling instead of the compute limit. An operational console built around a keyboard loop, non-blocking evidence, neighbour-frame browsing, an assignment board, and a submission builder answers C4. Within the special-session theme of AI systems for smart cities, AIODT is a multimodal system over a large urban video archive, of which the interface is one of four layers.

We make three system claims. First, we build a five-list rank-fusion retrieval architecture entirely from open-weight models, taking the heavy vision-language and speech channels and the reranker from the Qwen family and the dense and text encoders from PE-Core and bge-m3. We self-host it on two data-center-class GPU sessions, so that no closed model sits in the retrieval or serving core (Section III). A separate, swappable language model handles the Vietnamese-to-English bridge, and in the runs reported here a hosted API served it. Second, we design an operational layer around speed under pressure, with a keyboard loop, non-blocking evidence, and multi-operator collaboration through an assignment board (Section IV). Third, we report an operational evaluation on the official scorer that covers the infrastructure ceiling and the latency cost as well as the score (Section V).

II. Related Work

Two lines of research position the system layer of this paper, and a third positions its target deployment domains.

Interactive competition video retrieval. PicHunter [3] framed target search as Bayesian relevance feedback, Zahálka and Worring [4] cast large-collection exploration as a human–model dialogue, SOMHunter [5] organized that loop over a self-organizing map, and Lokoč et al. [6] introduced interaction logging at the Video Browser Showdown and concluded from those logs that user-centric interfaces are still required to reach specific content, which is the speed-under-pressure question our operational layer targets. Two recent Showdown entries set the interaction bar against which the console is measured: Vibro [7], winner of the 2022 edition, which pairs text-to-image and image-to-image search with browsing over 2D sorted maps, and VISIONE [8], runner-up in 2023, which integrates free-text, spatial colour and object, similarity, and temporal search. Ma and Ngo [9] framed interactive moment retrieval over a video corpus as reinforcement learning with human feedback, and their robust relevance feedback for known-item search remains active [10]; the Showdown result reports [1], [2] establish the human in the loop as a constituent component.

On the Vietnamese side, recent AIC systems converge on one multi-channel recipe: HelioSearch [11] pairs TransNetV2 [12] keyframes with dense embedding, OCR, ASR, detection, and LLM query expansion, and contemporary systems from the same competition share that recipe with queries translated into English before embedding: cascade embedding–rerank [13], LLandMark [14], MERVIN [15], and U-CESE [16]. A second cluster is closer still to our task: AViSearch [17] pairs query enhancement with optimized keyframe selection, MADTempo [18] formalizes interactive multi-event temporal retrieval, and two moment-retrieval systems add global reranking with bidirectional temporal search [19] and multi-granularity temporal reranking [20]. Fuse-plus-rerank is the regional consensus.

These works report system descriptions and standings, and the ones nearest our task build on closed models and a commercial image-search API [18] or report their results qualitatively [19], [20], since retrieval quality, not serving, is what those papers set out to measure. AIODT lies within this line architecturally, but we report the *operational layer*, an open-weight self-hosted serving core, and the infrastructure trade-offs these system papers rarely detail. In the adjacent direction of ultra-long video understanding, MERIT [21] keeps several keys per clip and defers refinement to temporal-neighbour search at inference time; the neighbour browsing and reranking of AIODT share that placement, and whether it holds under a frame-window scorer rather than answer accuracy is left to a corpus-level study.

Serving open multimodal models. Open vision-language families such as Qwen3-VL [22], together with the vLLM [23] and llama.cpp [24] serving stacks, make it possible to self-host the retrieval and serving core with no closed model, as we do here; the lighter query-side translation still calls a hosted LLM, kept swappable across providers. We contribute not these components but a two-GPU serving composition operated under real competition constraints, together with the infrastructure limits measured on it.

Target deployment domains. The strongest anchor in the surveyed literature is the AI City Challenge series [25], [26], whose warehouse spatial-intelligence track answers natural-language questions over camera networks alongside traffic-safety description and natural-language person search, the known-item family of the KIS channel. Natural-language retrieval for port and maritime logistics is nearly absent from that literature, where vision work stops at detection and tracking,

and it remains a hypothesis for this system, not a capability.

III. System Architecture

The task is fixed first. Given a Vietnamese natural-language query q over a video archive V, the system returns a ranked list of (video, frame) pairs, and a row is correct when its frame index falls inside the gold interval the organizer holds for q; the scoring rule over that list is given in Section V. Turning raw city footage into such a list takes four layers, shown in Fig. 1: ingestion-and-indexing, retrieval-and-fusion, model-serving, and operation. Each video is cut into shot keyframes with TransNetV2 [12], and each keyframe is indexed in parallel: a CLIP-family [27] dense backbone on the image, namely PE-Core [28], which is trained primarily on English; a caption written by Qwen3-VL-8B [22], which also produces the OCR text, embedded with bge-m3 [29] and indexed lexically; the OCR text and the ASR transcript from Qwen3-ASR-1.7B [30], indexed lexically; and detected objects. At query time five ranked lists are fused, the dense list from the English bridge string and the caption-embedding, caption, OCR, and ASR lists from the Vietnamese query, with weighted reciprocal-rank fusion [31] whose per-list weights are task priors, for known-item search 0.91 on the dense list, 0.19 on the caption and OCR lists, and 0.12 on ASR, and the top 30 of the fused list are reranked pointwise by a VLM, one judgement per query–keyframe pair. The object index is not fused; it is exposed to the operator as card evidence and as an optional entity filter. Query expansion into several variants and per-video diversification are implemented but were switched off in every scored run.

The Vietnamese query is rewritten and bridged into English before the dense channel is entered. One backbone limitation is recorded: the dense channel accepts 72 text tokens, and in our probes reordering the words of a query changed its top-ranked frame in most cases, so word order in the query matters as much as content. The running example of Section I shows why this layer is load-bearing: the three-sentence Vietnamese question is bridged to "Private spacecraft launch introduction, four astronauts wearing black suits, spacecraft in space studying polar aurora lights", a short string whose opening tokens carry the discriminative content.

The self-hosted serving layer is split across two sessions, both placed behind a public tunnel for the web console, on two A100-80GB on day one and on two RTX PRO 6000 Blackwell 96GB from day two: one serves the VQA model, Qwen3.8-27B [32] in the AWQ-INT4 release `cyankiwi/Qwen3.8-27B-AWQ-INT4` through vLLM on day one and Qwen3.8-Flash-Next [33] in the Q4_K_M GGUF release `AtomicChat/Qwen3.8-Flash-Next-GGUF` through llama.cpp from day two, so every VQA figure below is measured on 4-bit weights, and the other serves retrieval, namely the pointwise reranker on vLLM, PE-Core, and the text index. The same stack co-locates on a single 96GB GPU, but all three scored runs used the two-session split and that is what is reported. Every retrieval channel and the reranker run on open weights; the one component outside this boundary is the query-side language model for Vietnamese-to-English translation and query-side answering, served through a hosted API in our runs, a Gemini-class model by default with open-weight DeepSeek and Qwen fallbacks.

For an interactive loop that must answer in real time, the heaviest infrastructure constraint is not the GPU but the network layer, since any response longer than $100\,\mathrm{s}$ is cut by the tunnel; the concurrency decisions of Section V are balanced around that ceiling.

IV. Operational Console

The operational layer is the central system contribution, designed around one principle: the number of actions between seeing the right candidate and committing it to the submission is minimized, because under time pressure every mouse reach or context switch costs points.

Keyboard loop. In competition mode the entire browse-and-commit workflow is placed on the keyboard: arrow keys or h/j/k/l to move within the candidate grid, Enter to accept a frame into the submission tray, a dedicated shortcut to commit a question-answering row together with its answer, and number keys to switch among the task types.

Query-integrity aids. Because the Vietnamese-to-English query bridge carries the most weight and fails most often, it is exposed in the console, as Fig. 2 shows. An inline calculator converts a spoken timestamp in seconds into the exact frame index at the clip's frame rate. This removes a manual arithmetic step that, under time pressure, is a frequent source of off-by-frame errors.

Non-blocking evidence. The candidate grid is drawn first from the fast path, while each card's evidence, namely neighbour images, OCR and ASR snippets, and detected objects, is loaded asynchronously and independently per card. Fast-path latency is measured and shown at the foot of the screen, and is turned red past $1\,\mathrm{s}$.

Neighbour-frame browsing. A $\pm$context filmstrip around the frame is opened from each result card, so that adjacent keyframes can be stepped through until the exact moment is confirmed, and the committed frame can be swapped for a neighbour by one routine that keeps the result grid and the draft basket consistent. Each frame also carries a deep link to its exact source-video second for fast verification.

Distributed collaboration. A question set is split among multiple operators through an assignment board on which each question is locked, results are committed back, and the who-holds-what and who-submitted state is updated live. Competition-scale annotation proceeds without collisions.

Submission builder. The committed rows are collected in an always-visible submission tray that shows the exact CSV text to be submitted, and the organizer's row rules, namely the row-count cap, neighbour expansion, deduplication, and temporal ordering, are applied by a builder behind it before the final file is exported. A session clock marks the deadline and warns at time-out.

The evidence for this layer is bounded: one console decision, the submission builder's row-expansion rule, is isolated in Section V as a 0.4-point paired difference on one question set;

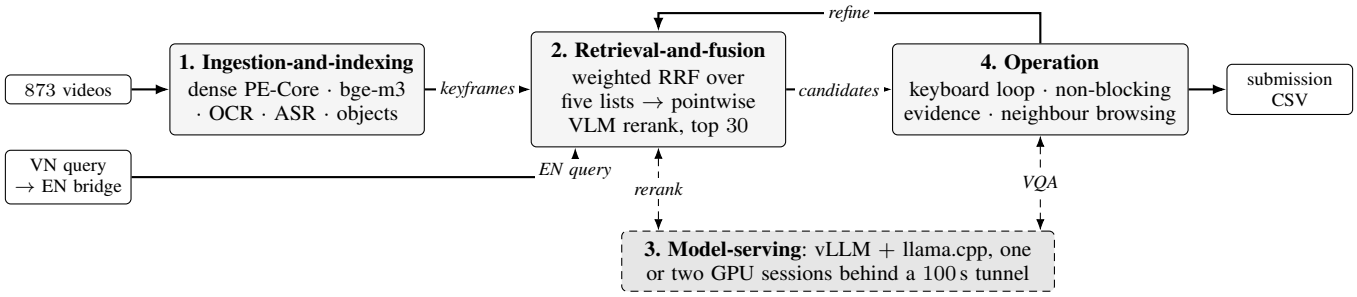

Fig. 1. AIODT dataflow. The archive is indexed once; a Vietnamese query enters retrieval through the English bridge, is answered by weighted rank fusion plus a pointwise VLM rerank, and reaches the operator, who either refines the query or commits the located frames as the answer.

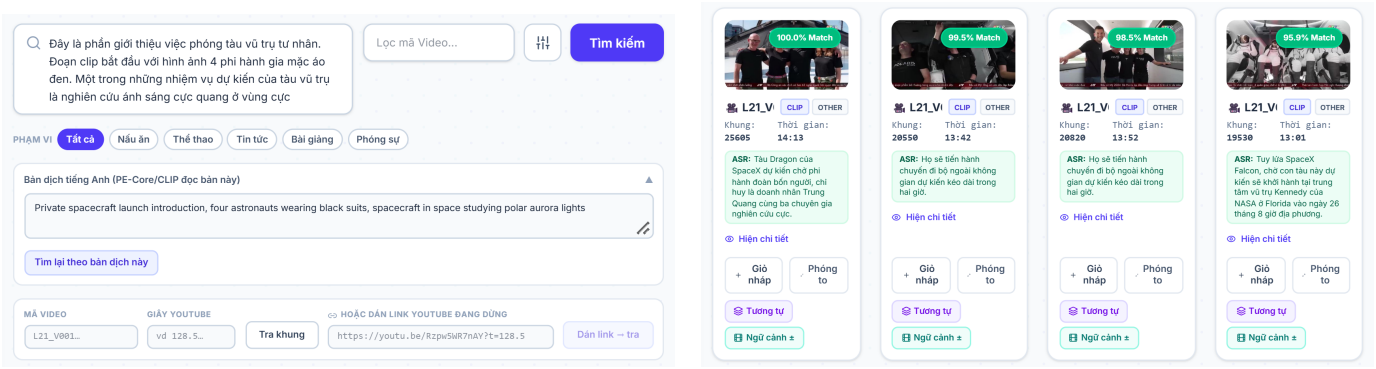

Fig. 2. The console for the running example. Left: the English string sent to the dense backbone is *exposed*, so a mistranslation can be corrected and re-searched in place, the operator's half of C1; the frame calculator sits below. Right: the first row of the ranked grid with per-card ASR evidence; the top card is a different clip of the same story and the target video appears from rank 2, the case neighbour-frame browsing exists for.

how many actions the keyboard loop or the assignment board saves per question is *not* measured by a controlled user study.

V. Operational Evaluation

AIODT is evaluated on the organizer's own submission portal; no internal engine produces the numbers below. Table I states the deployed configuration and the operating points discussed below.

Scoring rule. Each question is answered by a ranked list of up to 100 rows, each scored in $[0, 1]$ by membership of its frame index in a gold interval fixed by the organizer, and the question's Final Score is the mean of $R@k = \max_{i \le k} \text{R-Score}(r_i)$ over $k \in \{1, 5, 20, 50, 100\}$. Rank 1 therefore carries only one fifth of a question and a row first appearing at rank 40 still earns two fifths, which is why a submission fills the list rather than staking everything on one confident guess. Question answering additionally requires a semantic match on the answer string, and moment tracking is the only task with partial credit, the fraction of the N requested moments hit, collapsing to zero if the video is wrong.

Question sets. Day one posed 25 questions, of which 20 were known-item search, 4 question answering, and 1 moment tracking; day two posed 30, split 19, 9, and 2; and day three posed 36, split 26, 8, and 2. Half of each set is scored by the public leaderboard, so the attainable maxima on the visible portions are 13, 15, and 18 points.

Result. A day counts only its best submission, so our official figures are 9.8 of 13 on day one, 11.6 of 15 on day two, and 15.8 of 18 on day three, which is 75.4%, 77.3%, and 87.8% of the attainable maximum. Two further submissions were measured on the same official ruler without counting toward a day, 9.4 on day one and 12.6 on day three, and each is used below only as the control half of a same-day pair. The final score awaits the hidden round at the last competition session. Only same-day submissions are directly comparable, so the two day-one figures form one pair and the two day-three figures another.

Ablation scope. The two same-day pairs examined below are the paper's ablation instances: one isolates a submission-builder rule, the other isolates the operator, and each pair holds the question set and its maximum fixed. Component-level ablations, removing a channel, retuning the fusion weights, or dropping the reranker, require the fully automatic configuration of the same archive, the subject of a companion study; none is claimed here.

Submission construction. The two day-one submissions isolate one console decision on the official ruler, because the same question set and the same maximum scored both, and they differ only in how the submission builder filled each row list. In the 9.8 submission the builder expanded each committed frame into 41 uniformly spaced rows covering the operator's bracketed interval, whereas the 9.4 submission replaced that expansion with 10 hand-picked frames. Uniform coverage won

TABLE I
DEPLOYED CONFIGURATION AND MEASURED OPERATING POINTS; THE BINDING CONSTRAINT ON HOW FAST AN OPERATOR GETS AN ANSWER IS THE 100 S INGRESS CEILING, NOT THE GPUS

Configuration	
Corpus	AIC-2026, 873 videos
Shot segmentation	TransNetV2 [12]
Dense backbone	PE-Core [28], 72-token text context
Caption embedding	bge-m3 [29]
OCR / caption, index time	Qwen3-VL-8B [22]
Pointwise rerank	Qwen3-VL-Reranker-8B [34], vLLM
VQA, day 1	Qwen3.8-27B [32], AWQ-INT4
VQA, days 2–3	Qwen3.8-Flash-Next [33], Q4_K_M
ASR	Qwen3-ASR-1.7B [30]
Query bridge	hosted LLM, swappable
Serving back-ends	vLLM [23] ‖ llama.cpp [24]
Hardware, day 1	2× A100-80GB
Hardware, days 2–3	2× RTX PRO 6000 96GB
Measured operating points	
Fast path, retrieval → grid	< 1 s
VQA pass, 27B, 50 frames	47 s @ 1280×720, image share 15 s
Same pass @ long edge 640	-28%
Prefill share, 27B, batch 30	$\approx 94\%$ of a VQA turn
VQA turn, Flash-Next, A100	3.0 s
Fan-out, Flash-Next, 96GB	0.8 s per frame over 8 frames
Response ceiling, first byte	100 s
Best client concurrency	1 thread per GPU
Official score, public half of the set	
Day 1, 25 questions, max 13	9.8, best of 2
Day 2, 30 questions, max 15	11.6
Day 3, 36 questions, max 18	15.8, best of 2

by 0.4 points. The scoring rule explains the direction: coverage of an uncertain interval outranks one confident guess, and the neighbour-frame browsing and the expansion rule exist for this reason.

What the operator loop is worth. The two day-three submissions form a second pair on one question set and one maximum, and, unlike the day-one pair, they isolate the operator. We packed the 12.6 submission directly from the console's ranked output, whereas the 15.8 submission followed an in-console verification pass in which the operator re-scoped 8 of the 36 questions to a different video and promoted a verified frame to rank 1. Compared row by row, the two agree on the video for 28 of 36 questions but on the frame, within ± 5, for only 10, so the eight video-level changes are the largest candidate source of the $+3.2$-point gain, $+25\%$ in relative terms; a per-question decomposition awaits the released gold intervals. This pair is the closest we come to quantifying the operational layer, and it indicates that the loop earns its cost where the wrong video is picked by retrieval rather than where a nearby frame is picked. Both pairs are single paired observations, not controlled studies; only the direction is claimed.

Visual inference cost. A visual question-answering pass over the 50 reranked candidates costs 47 s of wall time on the day-one back-end, of which 15 s is the price of the pixels, and a single turn on the llama.cpp back-end of days two and three costs 3.0 s, or 0.8 s per frame when eight frames are fanned out. On vLLM a VQA turn is dominated by prefill, with only a small share spent on decode, so decode-side accelerations such as speculative decoding and MTP barely help; the pixels account for 15 s of the 47, so image size is the lever. These figures are at the competition setting, full-size 1280×720 images. In a post-competition measurement the long edge was lowered to 640, giving 640×360, shared across VQA, caption, and rerank, and 28% of a VQA turn's latency was cut with no loss on a face-counting bed; the rerank quality at 640 has *not* been measured separately.

Infrastructure ceiling and concurrency. The hard ceiling is the 100 s after which the free Cloudflare Quick Tunnel used for the competition returns HTTP 524, whatever the server is still doing. It is measured on the first byte, so a 102.8 s VQA turn still returned once bytes were emitted early. A single client thread per GPU is the best operating point; five threads raise batch throughput but lengthen every request toward the ceiling. The limit belongs to the tunnel tier, not to the models or the GPUs, so a direct ingress or a paid tier removes it.

Every score reported here was achieved by a human-machine pair on the public 50% portion of one corpus at one scale, with the hidden round excluded; it is not a decoupled measure of automatic capability, and a direct automatic baseline on the same official scorer is measured in the companion study on the fully automatic configuration of the same archive. The day-three pair bounds what the operator adds on top of the console's own output but does not separate the operator from the console. The latency figures come from the live system; no controlled load study has been run.

Data handling. The corpus is broadcast news and public urban footage, and no person identification is performed. Frames never leave the self-hosted back-ends; only the query text and the answer string pass through the hosted language model, which one setting repoints to the local vLLM instance. A deployment on private camera networks would also need a data-protection review that this work does not attempt.

VI. CONCLUSION

We presented AIODT, a self-hosted human-in-the-loop multimodal retrieval system for Vietnamese smart-city video archives, whose central contribution is the operational layer: a console optimized for speed under pressure, distributed collaboration, and an open-weight two-GPU serving composition. Across three competition days, each counting only its best submission, we reached 75 to 88% of the attainable maximum on the publicly revealed half of each question set, with the final score still pending the hidden round, at a sub-1 s fast path under the 100 s ceiling of the free tunnel. The visual-prefill cost, the tunnel ceiling, and the absence of a controlled user study are reported alongside capability. One paired day-three observation puts a first number on the operational layer: an in-console verification pass over the console's own ranked output changed the video on 8 of 36 questions and moved the score from 12.6 to 15.8.

Two directions follow. Within the operational layer we will widen the small-image path already measured at 640, worth -28% latency with rerank quality still unverified, run

a controlled study to replace this single pair with a measured gain, and take the console to the Video Browser Showdown, whose ad-hoc and conversational known-item tasks it already carries as practice modes. The larger question is what remains when the operator is removed, which an unattended city-scale deployment would require. The 8 video changes are the cases on which, in the operator's judgement, the console's own ranking had failed, so the operational gain doubles as a lower bound on the automatic path's residual error; locating that error, in the dense backbone, the fusion weights, or the reranker, is a measurement problem and the subject of a companion study on the fully automatic configuration of the same archive. The console, the indexing pipeline, and the serving notebooks will be released upon acceptance.

REFERENCES

[1] L. Rossetto, K. Schoeffmann, C. Gurrin, J. Lokoč, and W. Bailer, "Results of the 2024 video browser showdown," 2024, doi:10.48550/arXiv.2502.15683.

[2] L. Rossetto, K. Schoeffmann, C. Gurrin, J. Lokoč, and W. Bailer, "Results of the 2025 video browser showdown," 2025, doi:10.48550/arXiv.2509.12000.

[3] I. J. Cox, M. L. Miller, T. P. Minka, T. V. Papathomas, and P. N. Yianilos, "The Bayesian image retrieval system, PicHunter: Theory, implementation, and psychophysical experiments," *IEEE Trans. Image Process.*, vol. 9, no. 1, pp. 20–37, 2000, doi:10.1109/83.817596.

[4] J. Zahálka and M. Worring, "Towards interactive, intelligent, and integrated multimedia analytics," in *IEEE VAST*, 2014, pp. 3–12, doi:10.1109/VAST.2014.7042476.

[5] M. Kratochvíl, F. Mejzlík, P. Veselý, T. Souček, and J. Lokoč, "SOMHunter: Lightweight video search system with SOM-guided relevance feedback," in *ACM Multimedia*, 2020, pp. 4481–4484, doi:10.1145/3394171.3414542.

[6] J. Lokoč, G. Kovalčík, B. Münzer, K. Schöffmann, W. Bailer, R. Gasser, S. Vrochidis, P. A. Nguyen, S. Rujikietgumjorn, and K. U. Barthel, "Interactive search or sequential browsing? A detailed analysis of the video browser showdown 2018," *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 15, no. 1, pp. 29:1–29:26, 2019, doi:10.1145/3295663.

[7] K. Schall, N. Hezel, K. Jung, and K. U. Barthel, "Vibro: Video browsing with semantic and visual image embeddings," in *MultiMedia Modeling (MMM), LNCS*, 2023, pp. 665–670, doi:10.1007/978-3-031-27077-2_56.

[8] G. Amato, P. Bolettieri, F. Carrara, F. Falchi, C. Gennaro, N. Messina, L. Vadicamo, and C. Vairo, "VISIONE: A large-scale video retrieval system with advanced search functionalities," in *Proc. ACM Int. Conf. Multimedia Retrieval (ICMR)*, 2023, pp. 649–653, doi:10.1145/3591106.3592226.

[9] Z. Ma and C.-W. Ngo, "Interactive video corpus moment retrieval using reinforcement learning," in *ACM Multimedia*, 2022, pp. 296–306, doi:10.1145/3503161.3548277.

[10] Z. Ma and C.-W. Ngo, "Robust relevance feedback for interactive known-item video search," in *ACM ICMR*, 2025, pp. 982–990, arXiv:2505.15128; doi:10.1145/3731715.3733422.

[11] C. T. Vi, N. T. Luan, T. G. Nghi, M. N. P. Minh, N. D. Hoang Son, N. H. Quyen, and P. T. Duy, "HelioSearch: A multimodal video retrieval framework with LLM-driven query expansion and hybrid filtering," in *Information and Communication Technology (SOICT 2025)*, ser. CCIS. Springer, 2026, pp. 139–149, doi:10.1007/978-981-92-2584-2_11.

[12] T. Souček and J. Lokoč, "TransNet V2: An effective deep network architecture for fast shot transition detection," in *ACM Multimedia*, 2024, pp. 11 218–11 221, arXiv:2008.04838; doi:10.1145/3664647.3685517.

[13] T. T. Le Ngo, H. P. Ha, D. T. Nguyen Dang, M. T. Nguyen Le, and T. A. Nguyen Nhu, "Unified interactive multimodal moment retrieval via cascaded embedding-reranking and temporal-aware score fusion," 2025, arXiv:2512.12935, accepted at an AAAI 2026 workshop; doi:10.48550/arXiv.2512.12935.

[14] M.-C. Phung, T.-B. Le, C.-T. Tran-Thi, T.-D. Nguyen-Thi, V.-H. Dao, and T.-A. Nguyen-Nhu, "LLandMark: A multi-agent framework for landmark-aware multimodal interactive video retrieval," 2026, arXiv:2603.02888, accepted at the AAAI 2026 Workshop on New Frontiers in Information Retrieval.

[15] A.-T. Pham-Nguyen, T.-D. Le-Duc, A.-D. Le, and T.-H. Truong-Le, "MERVIN: A unified framework for multimodal event retrieval in Vietnamese news videos," in *Information and Communication Technology (SOICT 2025)*, ser. CCIS. Springer, 2026, pp. 204–215, arXiv:2605.16120; doi:10.1007/978-981-92-2597-2_16.

[16] D.-N. Le, H.-P. Nguyen, T.-D. Lam, M.-N. Dang, and M.-H. Le, "U-CESE: Unified CLIP-based event search engine for AI challenge HCMC 2025," in *Information and Communication Technology (SOICT 2025)*, ser. CCIS. Springer, 2026, pp. 489–503, arXiv:2605.23274; doi:10.1007/978-981-92-2590-3_40.

[17] N. H. H. Long, T. T. C. Giang, T. N. C. Nguyen, P. H. Phuoc, D. H. Phat, and T.-H. Nguyen, "AViSearch: A multimodal video event retrieval system via query enhancement and optimized keyframes," in *Information and Communication Technology (SOICT 2024)*, ser. CCIS. Springer, 2025, pp. 219–232, doi:10.1007/978-981-96-4291-5_18.

[18] H.-A. Vu, V.-K. Mai, T.-T. Nguyen, Q.-D. Dam, T.-H. Nguyen, and T.-H. Le, "MADTempo: An interactive system for multi-event temporal video retrieval with query augmentation," 2025, doi:10.48550/arXiv.2512.12929.

[19] T.-A. Nguyen-Nhu, H.-L. Tran, N.-K. Le, M.-N. Nguyen, T.-H. Nguyen *et al.*, "A lightweight moment retrieval system with global re-ranking and robust adaptive bidirectional temporal search," in *CVPR Workshops*, 2025, pp. 3708–3718, arXiv:2504.09298; doi:10.1109/CVPRW67362.2025.00356.

[20] H.-L. Tran, T.-A. Nguyen-Nhu, H.-P. Phan-Nguyen, T.-H. Nguyen *et al.*, "Towards efficient and robust moment retrieval system: A unified framework for multi-granularity models and temporal reranking," in *CVPR Workshops*, 2025, pp. 3719–3729, arXiv:2504.08384; doi:10.1109/CVPRW67362.2025.00357.

[21] Y. Choi, Y. Yoo, J.-Y. Lee, H. Kang, and S. J. Kim, "Keep it simple: Multi-key episodic memory retrieval for ultra-long video understanding," in *ECCV*, 2026, arXiv:2608.07663, accepted as oral; proceedings DOI pending; doi:10.48550/arXiv.2608.07663.

[22] Qwen Team, "Qwen3-VL technical report," 2025, doi:10.48550/arXiv.2511.21631.

[23] W. Kwon *et al.*, "Efficient memory management for large language model serving with PagedAttention," in *SOSP*, 2023, pp. 611–626, doi:10.1145/3600006.3613165.

[24] G. Gerganov *et al.*, "llama.cpp," https://github.com/ggml-org/llama.cpp, 2023.

[25] Z. Tang, S. Wang, D. C. Anastasiu, M.-C. Chang, A. Sharma *et al.*, "The 9th AI city challenge," in *ICCV Workshops*, 2025, arXiv:2508.13564; doi:10.1109/ICCVW69036.2025.00579.

[26] Z. Tang, S. Wang, D. C. Anastasiu, M.-C. Chang, A. Sharma *et al.*, "The 10th AI city challenge," 2026, arXiv:2608.17044, workshop in conjunction with ECCV 2026; doi:10.48550/arXiv.2608.17044.

[27] A. Radford *et al.*, "Learning transferable visual models from natural language supervision," in *ICML*, ser. PMLR, vol. 139, 2021, pp. 8748–8763, doi:10.48550/arXiv.2103.00020.

[28] D. Bolya, P.-Y. Huang, P. Sun, J. H. Cho, A. Madotto *et al.*, "Perception encoder: The best visual embeddings are not at the output of the network," in *NeurIPS*, 2025, pp. 67 743–67 796, arXiv:2504.13181; doi:10.52202/085713-2036.

[29] J. Chen, S. Xiao, P. Zhang, K. Luo, D. Lian, and Z. Liu, "M3-Embedding: Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation," in *Findings of ACL*, 2024, pp. 2318–2335, arXiv:2402.03216; doi:10.18653/v1/2024.findings-acl.137.

[30] Qwen Team, "Qwen3-ASR technical report," 2026, doi:10.48550/arXiv.2601.21337.

[31] G. V. Cormack, C. L. A. Clarke, and S. Büttcher, "Reciprocal rank fusion outperforms Condorcet and individual rank learning methods," in *Proc. ACM SIGIR*, 2009, pp. 758–759, doi:10.1145/1571941.1572114.

[32] Qwen Team, "Qwen3.8-Max: A new bar for coding and cowork," Aug. 2026, https://qwen.ai/blog?id=qwen3.8; model card https://huggingface.co/Qwen/Qwen3.8-27B.

[33] Qwen Team, "On the design of Qwen3.8-Next architecture: Evaluation, efficiency, and training stability," 2026, doi:10.48550/arXiv.2608.30320.

[34] M. Li, Y. Zhang, D. Long, K. Chen, S. Song, S. Bai, Z. Yang, P. Xie, A. Yang, D. Liu, J. Zhou, and J. Lin, "Qwen3-VL-Embedding and Qwen3-VL-Reranker: A unified framework for state-of-the-art multimodal retrieval and ranking," 2026, doi:10.48550/arXiv.2601.04720.