BEPA: AGENT-GUIDED EVIDENCE POLICY ADAPTATION FOR VIDEO QUESTION ANSWERING

Anonymous authors
Paper under double-blind review

ABSTRACT

Under a limited frame budget, question-relevant observations may remain compatible with multiple answers, while shared scene content obscures discriminative cues. We propose BEPA (Agent-Guided Evidence Policy Adaptation) to organize evidence around inter-option distinguishability and signal–background separability with a frozen video answerer. Before answering, CORE (Contrastive Option-guided Refinement of Evidence) suppresses shared option responses and refines a few BOLT fill frames. In an independent post-answer setting, a text agent turns prediction disagreement into visual queries for one additional observation. On 4,703 VUDG questions across 11 domains with VideoLLaMA3-7B, exploratory evaluation yields 75.82% domain-averaged accuracy (D-Avg) for CORE, versus 75.16% for matched BOLT, using one answer and at most 32 frames. Independent agent-guided acquisition reaches 77.26% D-Avg at 2.52 video answers per question on average.

1 INTRODUCTION

Video question answering requires identifying question-relevant objects, attributes, and interactions in dynamic scenes. However, visual context and computational budgets often restrict the answerer to a subset of video frames, so decisive states or brief interactions may be omitted. Organizing limited visual evidence is therefore important for improving video question answering without updating model parameters.

Prior work improves evidence acquisition in several ways. BOLT samples frames by question–frame relevance (Liu et al., 2025). WFS-SB organizes clips around semantic boundaries while balancing relevance and diversity (Chen et al., 2026). Active video perception further acquires evidence as needed through planning, observation, and feedback (Wang et al., 2025; Zou et al., 2026).

1.1 CHALLENGES

Despite advances in content localization, temporal coverage, and active acquisition, organizing evidence that distinguishes candidate answers under a limited budget still faces two challenges.

(1) Relevant observations may not distinguish answers. Observation relevance is not equivalent to its value for distinguishing candidate answers. In multiple-choice video question answering, the question and options often share the same scene, while the answer depends on finer distinctions in an object's state, an action's endpoint, or a temporal relation. For example, seeing a person holding an object is highly relevant but does not reveal whether it is eventually placed on a table, in a drawer, or in a cabinet; the decisive observation is where the object finally comes to rest. Under a fixed budget, repeated coverage of the shared scene can displace observations that capture such distinctions.

(2) Shared content can obscure discriminative cues. Evidence acquisition should emphasize where candidate answers differ rather than continue to reward repeated shared content. Similar objects, redundant backgrounds, and scene descriptions shared across the question and options can all produce high matching responses to the same frames, while simply appending each option to the question does not ensure that retrieval scores emphasize these differences.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

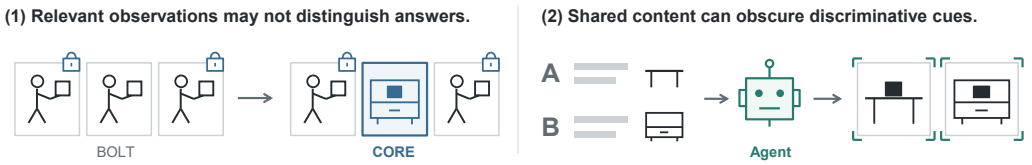

Figure 1: Challenges and BEPA responses, illustrated by asking where an object is placed. Left: repeated frames of a person holding the object are relevant to both "on the table" and "in the drawer," but do not resolve its destination. Locks mark the original BOLT anchors that CORE preserves; the blue frame illustrates an option-contrastive fill-frame replacement that reveals the destination without increasing the 32-frame budget. Right: options A/B share scene descriptions (gray strokes) but differ in destination (table/drawer icons). After BOLT/Uniform disagreement, the text-only Agent (robot) converts this distinction into two visual queries. Teal brackets mark the query-directed evidence targets, retrieved as one additional observation for the frozen video answerer. The two settings are evaluated independently; all scenes and selections are schematic.

Our Key Insight. Evidence organization should add a candidate-discriminative objective on top of relevance and coverage. This objective has two complementary aspects: **inter-option distinguishability** emphasizes object, attribute, and relation cues that separate candidate answers, while **signal–background separability** makes these differences more prominent in frame scores and retrieval queries, reducing the influence of shared scenes and repeated wording. Discriminative selection should therefore complement relevance-based strategies rather than replace contextual evidence.

1.2 Contributions

We instantiate this view as **BEPA (Agent-Guided Evidence Policy Adaptation)**, an evidence-policy adaptation framework for frozen video answerers. BEPA exploits discriminative requirements at two decision points: candidate options provide semantic differences before answering, while predictions from different observations reveal more specific disputes afterward. These signals guide evidence acquisition before and after prediction, respectively, and we evaluate the two settings independently. Figure 1 illustrates their relationship. Our contributions are:

- **CORE: Contrastive Option-guided Refinement of Evidence.** Building on BOLT's relevance-based sampling, CORE subtracts the per-frame mean option-matching response and constructs an option-coverage objective from relative responses. It protects the original sampling anchors and replaces at most four fill frames, using one answer and at most 32 frames.

- **Agent-Guided Targeted Evidence Acquisition.** On an independent base of BOLT, Uniform, and local WFS-CLIP, a text-only Qwen planner converts BOLT/Uniform disagreement into visual queries about disputed objects, attributes, or actions for one additional observation. The frozen video model answers from retrieved frames, fixed probability fusion returns the final answer, and answer-preserving execution omits calls that cannot change the fixed three-view fusion answer.

- **Budget-Constrained Evaluation and Utility Analysis.** We separately evaluate single-observation refinement and post-disagreement acquisition across VUDG's 11 domains. Published benchmark results provide context, while matched single-view controls measure answering utility. Question-level comparisons distinguish answer changes, error repairs, and harms to correct answers. C2 results report accuracy alongside logical video-answering and planning costs.

2 Related Work

2.1 From Relevant Content to Temporal Coverage

BOLT localizes relevant content by sampling question–frame similarities, while AKS jointly considers relevance and video coverage (Liu et al., 2025; Tang et al., 2025). WFS-SB detects semantic boundaries from multiscale relevance changes and balances relevance with diversity under clip budgets (Chen et al., 2026). GCR integrates timestamped text, visual anchors, and context, revisiting omitted intervals to replace weaker evidence (Wei et al., 2026). We further focus on distinguishing candidate answers by complementing question relevance with relative responses across options.

2.2 From Question-Conditioned Selection to Feedback-Driven Acquisition

DIG chooses between coverage and localization based on query type (Li et al., 2026). AutoSkill learns frame-selection skills from labeled examples and selects video inputs according to question semantics (Hu et al., 2026). Feedback-driven methods instead adapt acquisition to current observations: A.I.R. iteratively analyzes small batches of candidate frames, while AVP acquires timestamped evidence through planning, observation, and reflection (Zou et al., 2026; Wang et al., 2025). BEPA uses competing predictions across observations to determine what candidate distinctions should be checked next.

2.3 Cross-Domain Evaluation and Observation Utility

STDN studies video domain generalization through diversified spatiotemporal perception, and ARGO1M examines action recognition across scenarios and locations (Lin et al., 2023; Plizzari et al., 2023). VUDG provides video question answering across 11 visual domains within a shared activity scope (Wang et al., 2026). We evaluate evidence policies for a frozen answerer in this setting, combining domain-averaged accuracy with question-level repairs and harms. Cross-domain comparison additionally requires common semantic support, which remains to be established.

3 Methodology

3.1 Organizing Observations Around Candidate Distinctions

BEPA uses option semantics before answering and prediction disagreement after answering to guide evidence acquisition under a limited observation budget. CORE refines BOLT observations using relative matching responses across options; agent-guided acquisition translates competing predictions from different observations into visual queries for one additional observation.

Given video $V = \{(x_i, t_i)\}_{i=1}^{T}$, question q, and six options $\mathcal{A}$, t_i is the source frame's actual timestamp. A selection policy chooses source-frame set S, ordered into observation $E(S)$. The frozen answerer F_θ produces option logits ℓ_S, giving

$$p_S(a) = \mathrm{softmax}(\ell_S)_a, \qquad \hat{a}_S = \arg\max_{a \in \mathcal{A}} p_S(a). \tag{1}$$

As shown in Figure 2, CORE (C1) and agent-guided acquisition (C2) are evaluated independently. Cross-domain analysis (C3) remains pending.

3.2 CORE: Contrastive Option-guided Refinement of Evidence

CORE builds on BOLT's relevance-based sampling, using option-contrastive scores and a coverage objective to select a few fill-frame swaps that seek cues distinguishing the options within a fixed frame budget (Figure 3).

Suppressing shared responses. CORE reuses CLIP visual features from BOLT's candidate pool $\mathcal{C}$ and encodes six option-text-plus-stem queries without option letters. Let L_{ia} be the CLIP matching logit between candidate frame i and the query for option a. Shared scene content can match

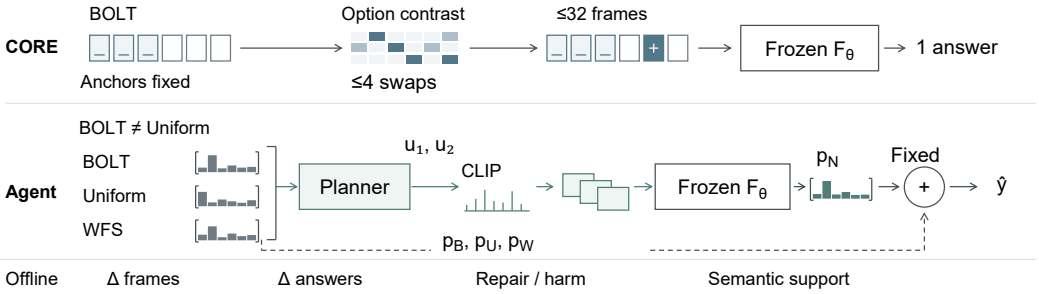

Figure 2: BEPA overview. CORE refines BOLT observations for one answer; the independent agent path acquires evidence after disagreement. Dashed lines carry base probabilities to fixed four-way fusion. The bottom row denotes planned offline utility and semantic-support checks, not completed C3 findings. All visual elements are schematic.

several queries, so we subtract the per-frame option mean and retain positive deviations as relative responses:

$$e_{ia} = \max\left(L_{ia} - \frac{1}{6}\sum_{b\in\mathcal{A}} L_{ib}, 0\right), \qquad u_{ia} = \frac{e_{ia}}{\max_j e_{ja}}\left(\frac{1}{2} + \frac{1}{2}w_i\right). \tag{2}$$

Here, $\max_j$ ranges over candidate pool $\mathcal{C}$. Per-option peak normalization equalizes response scales, with all-zero columns mapped to zero. The stem score w_i is obtained by within-question min–max normalization followed by cubing, retaining BOLT's stem-relevance preference. These relative scores guide retrieval; they are not option-correctness probabilities. Duplicate source indices share one slot with per-option maximum utility.

Covering candidate distinctions. To avoid rewarding redundant frames, coverage averages each option's maximum utility:

$$J(S) = \frac{1}{6}\sum_{a\in\mathcal{A}} \max_{i\in S} u_{ia}. \tag{3}$$

The per-option maximum prevents repeated accumulation of already covered responses: an added frame contributes by improving the observation's option coverage. This surrogate weights all six options equally and is computed directly from CLIP matching scores.

Preserving context through bounded swaps. Option coverage guides which cues to add, while BOLT anchors retain question-relevant context. Let S_0 be BOLT's original input and $P \subseteq S_0$ all unique power-3 inverse-transform-sampling hits. CORE fixes P and adjusts only the original fill set $S_0 \setminus P$. At each step, it replaces one unmodified fill frame with a not-yet-added candidate outside S_0, choosing the eligible swap with the greatest gain in J, for at most four swaps. Newly added frames remain fixed. For eligible 32-frame inputs,

$$P \subseteq S, \qquad |S| = 32, \qquad |S \cap S_0| \geq 28. \tag{4}$$

Final frames are ordered by actual timestamps and passed to the frozen answerer once. Additional encoding consists of six CLIP text queries.

In the implementation, gains must exceed 10^{-6}, and source-index order resolves ties deterministically. Short candidate pools, constant stem relevance, 32 protected anchors, or no eligible initial swap retain the original input. This 32-frame configuration corresponds to the sealed implementation `BOLT-OCS32`.

3.3 From Prediction Disagreement to Targeted Acquisition

Competing predictions from different observations provide a concrete target for acquisition after answering. C2 defines its base answer by equally fusing probabilities from BOLT, Uniform, and local WFS-CLIP observations B, U, W, each answered by the same frozen model with at most 32

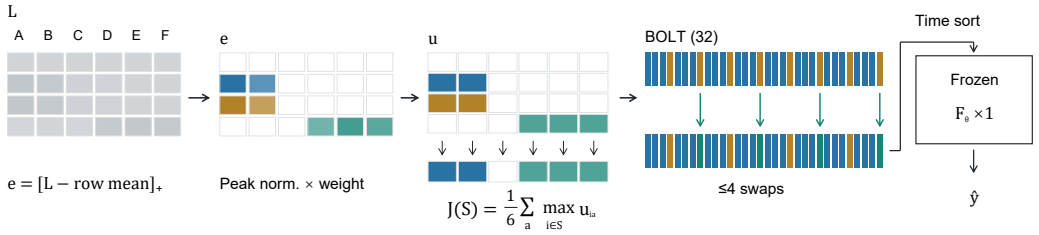

Figure 3: CORE. Option-contrastive coverage guides up to four BOLT fill-frame swaps before one answer. Blue, ochre, and teal denote protected anchors, original fills, and new frames; gray denotes shared responses. Matrices and frame sequences are schematic.

frames. Execution first obtains BOLT and Uniform predictions and uses their disagreement as the trigger:

$$g = \mathbf{1}[\hat{a}_B \neq \hat{a}_U], \tag{5}$$

When $g = 1$, C2 acquires WFS to complete the three-view summary, then triggers one targeted acquisition (Figure 4). When $g = 0$, no targeted acquisition is made: C2 returns the three-view base answer, omitting WFS only when the answer-preserving condition in Section 3.4 holds.

Turning a dispute into queries. A text-only Qwen agent receives question q, options $\mathcal{A}$, duration D, and summary H containing the three predictions, option probabilities, and observed timestamps. It proposes a fact to check that distinguishes competing interpretations, together with two visual queries:

$$\pi(q, \mathcal{A}, D, H) \longrightarrow (f, u_1, u_2). \tag{6}$$

The planning prompt asks queries u_1, u_2 to describe observable objects, attributes, or actions and cover competing interpretations where possible. Temporal questions prompt separate searches for the relevant events; object or attribute questions prompt searches for the object and its interaction context. The fact f records the verification purpose. Invalid outputs fall back to the original stem.

Acquiring a query-directed observation. To obtain temporal context around each query match, sparse CLIP matching locates one anchor per query, followed by sampling in its neighborhood. Given observed source frames $M = S_B \cup S_U \cup S_W$, retrieval prioritizes unseen frames, deduplicates and fills fixed quotas, then orders actual timestamps:

$$E_N = \mathcal{R}(V, u_1, u_2; M), \qquad |E_N| \leq 32. \tag{7}$$

The configuration uses two 12-frame windows of radius two seconds, plus eight global frames, with anchors separated by at least four seconds when feasible. Sparse features locate windows; actual sampling uses the source timeline. If no new source frame is available, the base answer is returned.

Answering from the additional frames. The answerer receives E_N, the original question, and original options; planning text acts only on retrieval. The new distribution p_N is fused equally with the base probabilities:

$$\hat{y} = \arg\max_{a \in \mathcal{A}} \frac{p_B(a) + p_U(a) + p_W(a) + p_N(a)}{4}. \tag{8}$$

Acquisition then ends. C2 uses at most one text-planning call and two to four video answers per question, with at most 32 frames per answer.

3.4 Answer-Preserving Budget Control

Targeted acquisition uses three base predictions, but for some questions the first two already determine the base option. We therefore omit WFS calls that cannot change the fixed three-view fusion option. Let $s = p_B + p_U$ and $a^* = \arg\max_a s(a)$, with leading margin

$$\Delta = s(a^*) - \max_{b \neq a^*} s(b). \tag{9}$$

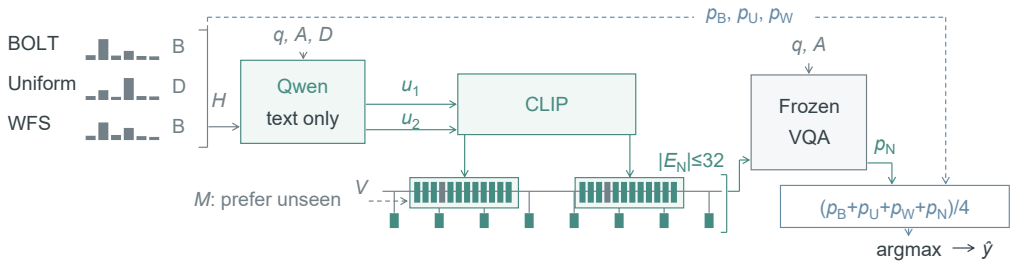

Figure 4: Targeted acquisition. BOLT–Uniform disagreement triggers two queries and one new observation. The new and three base probability distributions are averaged equally. Frames and predictions are schematic.

Any third distribution supplies at most one unit of relative advantage to a competitor, so

$$[s(a^*) + p_W(a^*)] - [s(b) + p_W(b)] \geq \Delta - 1. \tag{10}$$

If $\Delta > 1 + \varepsilon$, return a^* directly; otherwise acquire WFS and form the three-view base prediction, with $\varepsilon = 10^{-12}$. Questions with $g = 1$ then proceed to targeted acquisition in Section 3.3; those with $g = 0$ return the base answer. The omission rule preserves the base option of fixed three-view fusion. When BOLT and Uniform disagree, Δ cannot strictly exceed one. Early return therefore occurs only on non-triggered questions, and targeted acquisition always receives all three summaries.

3.5 From Observation Changes to Answering Utility

A matched comparison fixes the answering interface, budgets, and trigger population. Source-frame Jaccard measures observation differences. For condition c, reference c_0, and offline gold option y_i, define

$$\begin{aligned} D_i(c; c_0) &= \mathbf{1}[\hat{y}_i^c \neq \hat{y}_i^{c_0}], \\ U_i(c; c_0) &= \mathbf{1}[\hat{y}_i^c = y_i] - \mathbf{1}[\hat{y}_i^{c_0} = y_i]. \end{aligned} \tag{11}$$

Here $U_i = -1, 0, 1$ denotes harm, unchanged correctness, and repair. Repair and harm rates use reference errors and reference-correct answers, respectively. Aggregate utility is the 11-domain D-Avg difference; resampling clusters questions by source video. Third-view analysis compares $(p_B + p_U)/2$ with $(p_B + p_U + p_W)/3$ within agreement and disagreement groups; fourth-view utility is limited to its measured trigger population. C3 requires comparable question semantics and controls for question type, duration, candidate frames, visual tokens, and initial difficulty. Lexical similarity alone does not establish semantic equivalence.

4 Experiments

4.1 Experimental Setup

We evaluate 4,703 six-option VUDG questions across 11 domains (Wang et al., 2026), using frozen VideoLLaMA3-7B for answering, CLIP for retrieval, and Qwen3-4B-Instruct-2507 for C2 text planning (Zhang et al., 2025; Radford et al., 2021; Qwen Team, 2025). The primary metric is domain-equal accuracy, D-Avg $= (100/11) \sum_{d=1}^{11} c_d/n_d$, where c_d of n_d questions are correct in domain d. Correct counts are diagnostics, not a replacement metric.

C1 fixes one answer and at most 32 frames. CORE, BOLT32, and WFS-CLIP32 share the original BOLT flat prompt; Uniform32 retains its global-coverage description and is only a tool reference. C2 uses a separate BOLT/Uniform/WFS base, two to four video answers of at most 32 frames each, and at most one planner call. The settings are evaluated independently, not as a combined pipeline. Evaluation is development-exposed; reported 95% intervals use 2,000 domain-stratified video-bootstrap draws without multiplicity correction.

4.2 Comparison with Published VUDG Results

We report all 11 models in the original VUDG multiple-choice zero-shot table. Qwen2.5VL-7B is its best entry. Published and local results retain their distinct configurations.

Table 1: Published VUDG multiple-choice zero-shot references (Wang et al., 2026, Table 6). These are not local replications or budget-matched controls.

Published model	D-Avg (%)
Video-ChatGPT-7B	13.6
MiniGPT4-Video	13.8
VideoChat2-7B	14.0
Video-LLaVA-7B	23.5
VideoLLaMA2-7B	33.4
mPLUG-Owl3-7B	51.4
Video-CCAM-7B	51.5
Tarsier2-7B	62.8
GPT-4o (16 frames)	64.6
VideoLLaMA3-7B	65.1
Qwen2.5VL-7B (best in the original table)	**72.1**

CORE and C2 exceed the original table's best point estimate by 3.72 and 5.16 percentage points, respectively. These cross-configuration gaps are descriptive, not controlled SOTA or causal method gains: frame selection, processing, decoding, and C2 call budgets differ. Relative to published VideoLLaMA3-7B, the gaps are 10.72 and 12.16 points.

4.3 Fixed-Budget Refinement

This comparison tests whether refining BOLT fill frames while preserving its context improves answering without additional video calls or frames.

Table 2: C1: one answer, at most 32 frames. Uniform32 is not prompt-matched to CORE. Highlighting denotes point estimates, not significance.

Method	D-Avg (%)	Correct / 4,703
Uniform32 (tool reference)	74.237981	3506
BOLT32	75.157085	3540
WFS-CLIP32	75.276855	3549
CORE (C1, ours)	**75.820150**	**3569**

CORE improves over matched BOLT by 0.663 pp $[0.319, 0.980]$, repairing 46 answers and harming 17. Its 0.543-point difference from WFS-CLIP has interval $[-0.426, 1.446]$. Bounded refinement improves the matched BOLT result, while the difference from WFS remains inconclusive. CORE adds six CLIP text encodings, so its fixed video budget does not imply zero extra computation. The comparison supports the complete CORE configuration, not the individual effects of suppression, anchor protection, or swap budget. Uniform32's pooled accuracy, $3506/4703 = 74.548161\%$, differs from its domain-equal D-Avg. With the answerer, prompt, and video budget held fixed, the gain over BOLT supports improving evidence composition rather than increasing answering attempts; the remaining harms also show that refinement is not uniformly beneficial.

Observation change versus useful evidence. Observation changes alone do not certify useful evidence: in the separate BOLT/WFS comparison, 3,829 different-input questions produce 229 repairs, 220 harms, and 100 wrong-to-wrong changes; 874 identical-input questions show no answer changes. Thus 85.66% of different-input questions retain the same answer. Full-set WFS minus BOLT is $+0.120$ pp $[-0.843, 1.045]$. These counts describe a separate single-view comparison.

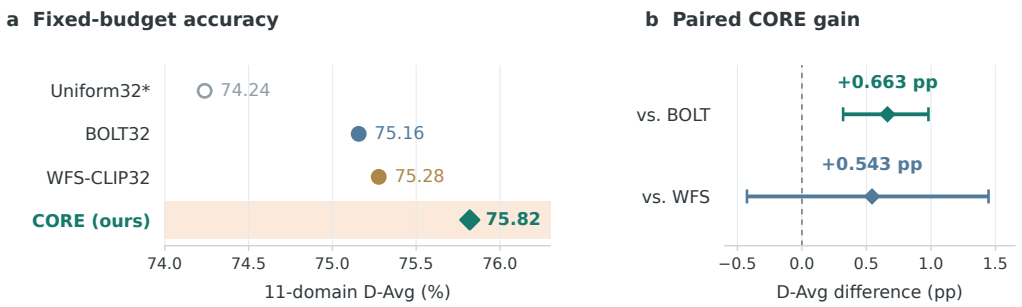

Figure 5: C1 at one answer and at most 32 frames. Left: full-set D-Avg; * marks Uniform32 as a non-prompt-matched tool reference. Right: paired CORE-minus-control differences with exploratory 95% video-bootstrap intervals. The WFS interval crosses zero.

The nearly balanced repairs and harms are consistent with the small aggregate difference. This contrasts with CORE's repair–harm balance and motivates evaluating evidence by its effect on answer correctness, rather than treating visual diversity as a proxy for utility.

4.4 Directed Acquisition and Execution Controls

Table 3 reports execution and acquisition configurations. Fixed-two and fixed-three fuse the base tools; answer-preserving execution omits calls that cannot change the fixed-three option. Low-margin routing uses fewer calls, while coverage, stem retrieval, and Qwen-directed verification provide alternative acquisition strategies.

Table 3: C2 execution and acquisition results on 4,703 questions across 11 domains. VQA and planner columns are mean logical calls per question.

Condition	D-Avg (%)	Correct	VQA	Planner
Low-margin logit fusion	74.668857	3526	1.288	0
Fixed BOLT + Uniform	74.712393	3529	2.000	0
Fixed three-view fusion	75.498280	3563	3.000	0
Answer-preserving execution	75.498280	3563	2.375	0
Additional global coverage	75.607044	3568	2.509	0
Additional stem-only retrieval	75.520198	3565	2.509	0
BEPA: Qwen-directed verification	**77.263073**	**3640**	2.518	0.151

Qwen reaches 77.263073% D-Avg with 3,640 correct answers. The run uses 11,129 base VQA calls and 712 additional VQA calls, totaling 11,841, plus 712 planning calls: 688 valid plans and 24 stem fallbacks. Its added observations comprise 22,784 unseen source frames and 7,338,144 visual tokens. Logical calls include cache reuse and do not measure online latency. Planning is selective, but a valid plan need not retrieve decisive evidence. The result characterizes the complete acquisition-and-fusion configuration.

5 Conclusion

BEPA adapts evidence through option-contrastive BOLT refinement and disagreement-guided acquisition. In separate evaluations, CORE improves single-answer accuracy over matched BOLT, and Qwen-directed verification reaches 77.263073% D-Avg. New evidence can help or harm answers, motivating joint assessment of answering utility and acquisition cost.

AI USE STATEMENT

We used generative AI solely to improve the grammar, wording, and readability of the manuscript originally written by the authors. The authors reviewed each revision to ensure that the technical meaning and conclusions were preserved. Generative AI was not used for any activity requiring mandatory disclosure; such activities were not applicable to this study. The authors take full responsibility for the final content.

ETHICS STATEMENT

This work studies evidence selection on the existing VUDG video QA benchmark. Pretrained models and source videos may carry biases and privacy risks; selecting additional evidence does not eliminate these risks or guarantee correct answers. High-stakes or surveillance use requires separate safeguards and validation. Source videos and case thumbnails remain subject to their original licenses and display permissions.

REPRODUCIBILITY STATEMENT

Sections 3 and 4.1 specify the policies and evaluation protocol. Appendix A records implementation settings, development limitations, and evidence provenance.

REFERENCES

Wang Chen, Yuhui Zeng, Yongdong Luo, Tianyu Xie, Luojun Lin, Jiayi Ji, Yan Zhang, and Xiawu Zheng. Wavelet-based Frame Selection by Detecting Semantic Boundary for Long Video Understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24052–24061, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Chen_Wavelet-based_Frame_Selection_by_Detecting_Semantic_Boundary_for_Long_Video_CVPR_2026_paper.html.

Jian Hu, Zixu Cheng, Da Li, Wei Li, Ziquan Liu, and Shaogang Gong. One Skill Does Not Fit All: Automatic Discovery and Taxonomy-Guided Routing of Frame-Selection Skills for Long-Video Question Answering. arXiv preprint arXiv:2609.12517, 2026. URL https://arxiv.org/abs/2609.12517.

Jialuo Li, Bin Li, Jiahao Li, and Yan Lu. Divide, then Ground: Adapting Frame Selection to Query Types for Long-Form Video Understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11369–11380, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Li_Divide_then_Ground_Adapting_Frame_Selection_to_Query_Types_for_CVPR_2026_paper.html.

Kun-Yu Lin, Jia-Run Du, Yipeng Gao, Jiaming Zhou, and Wei-Shi Zheng. Diversifying Spatial-Temporal Perception for Video Domain Generalization. In *Advances in Neural Information Processing Systems*, volume 36, 2023. doi: 10.52202/075280-2443. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/aef34c770664d06eabdfebc5d3d58a9c-Abstract-Conference.html.

Shuming Liu, Chen Zhao, Tianqi Xu, and Bernard Ghanem. BOLT: Boost Large Vision-Language Model Without Training for Long-form Video Understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3318–3327, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Liu_BOLT_Boost_Large_Vision-Language_Model_Without_Training_for_Long-form_Video_CVPR_2025_paper.html.

Chiara Plizzari, Toby Perrett, Barbara Caputo, and Dima Damen. What Can a Cook in Italy Teach a Mechanic in India? Action Recognition Generalisation Over Scenarios and Locations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 13656–13666, 2023. URL https://arxiv.org/abs/2306.08713.

Qwen Team. Qwen3-4B-Instruct-2507. Official model card, 2025. URL https://huggingface.co/Qwen/Qwen3-4B-Instruct-2507. Accessed September 24, 2026.

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. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pp. 8748–8763. PMLR, 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Xi Tang, Jihao Qiu, Lingxi Xie, Yunjie Tian, Jianbin Jiao, and Qixiang Ye. Adaptive Keyframe Sampling for Long Video Understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 29118–29128, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Tang_Adaptive_Keyframe_Sampling_for_Long_Video_Understanding_CVPR_2025_paper.html.

Ziyang Wang, Honglu Zhou, Shijie Wang, Junnan Li, Caiming Xiong, Silvio Savarese, Mohit Bansal, Michael S. Ryoo, and Juan Carlos Niebles. Active Video Perception: Iterative Evidence Seeking for Agentic Long Video Understanding. arXiv preprint arXiv:2512.05774, 2025. URL https://arxiv.org/abs/2512.05774. Version 2, June 2026.

Ziyi Wang, Zhi Gao, Boxuan Yu, Zirui Dai, Peiyao Wang, Yuxiang Song, Qingyuan Lu, Jin Chen, and Xinxiao Wu. VUDG: A Dataset for Video Understanding Domain Generalization. In *International Conference on Learning Representations (ICLR)*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/6aa3267e3a0f7c2c718036d08d482a51-Abstract-Conference.html.

Fan Wei, Siru Zhong, Runmin Dong, Miao Yang, Zhaoyang Luo, and Haohuan Fu. Ground, Cover, and Refine: Evidence-Centric Frame Selection for Long-Video Question Answering. arXiv preprint arXiv:2608.01660, 2026. URL https://arxiv.org/abs/2608.01660.

Boqiang Zhang, Kehan Li, Zesen Cheng, Zhiqiang Hu, Yuqian Yuan, Guanzheng Chen, Sicong Leng, Yuming Jiang, Hang Zhang, Xin Li, Peng Jin, Wenqi Zhang, Fan Wang, Lidong Bing, and Deli Zhao. VideoLLaMA 3: Frontier Multimodal Foundation Models for Image and Video Understanding. arXiv preprint arXiv:2501.13106, 2025. URL https://arxiv.org/abs/2501.13106.

Yuanhao Zou, Shengji Jin, Andong Deng, Youpeng Zhao, Jun Wang, and Chen Chen. A.I.R.: Enabling Adaptive, Iterative, and Reasoning-based Frame Selection For Video Question Answering. In *International Conference on Learning Representations (ICLR)*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/1332893b662f655660c9abdf793230cf-Abstract-Conference.html.

A SUPPLEMENTARY NOTES

Reproduction. C2 used an RTX 5090, PyTorch 2.8.0+cu128, Transformers 4.50.0 for answering/retrieval and 4.51.3 for planning, and Decord 0.6.0. The deterministic planner allowed at most 6,144 input and 256 output tokens. The sealed C1 environment, public planner/CLIP snapshot mappings, and complete processor settings remain to be documented.

Development scope. Answer-preserving execution was selected after the sequential experiment. The 120-question verification pilot failed its original threshold; the full evaluation retained the same policy and pilot questions. These development-exposed results are exploratory, and the existing data halves are not independent validation sets.

Evidence records. Scorer summaries, prediction manifests, and root seals identify the reported C1 and Qwen C2 result. Full artifact paths, hashes, implementation identities, and historical run records are retained in the working evidence records.