

GOHi: Goal-Oriented Highlight Extraction for Long-Form Egocentric Video

KAYLEE YAXUAN LI , University of Michigan, USA and Reality Labs Research, Meta, USA

KE SUN, University of Michigan, USA

KANG G. SHIN, University of Michigan, USA

ALANSON P. SAMPLE, University of Michigan, USA

AMY KARLSON, Reality Labs Research, Meta, USA

ERIC WHITMIRE, Reality Labs Research, Meta, USA

YIFAN WANG, Reality Labs Research, Meta, USA

Personal AI assistants are moving beyond isolated interactions toward continuous, context-aware assistance grounded in a user's experience. Providing such assistance requires episodic memory that can recall past events, interpret everyday activities, and support long-horizon reasoning. Egocentric video offers a natural record of these experiences, but meaningful actions are sparse within highly redundant footage, and the volume of visual input grows with recording duration. Processing raw video therefore incurs substantial computational and memory costs, may exceed model context limits, and dilutes salient evidence with irrelevant content. The central challenge is to transform continuous egocentric video into a compact episodic representation that preserves essential context while removing redundancy. This paper presents GOHi, a goal-oriented framework that transforms continuous egocentric video into a compact, temporally coherent sequence of keyframes. The central insight is that the steps required to achieve a user's goal provide natural semantic and temporal structure for identifying salient moments. A pretrained vision-language encoder aligns video frames with natural-language descriptions of these steps to identify goal-relevant visual evidence. The resulting similarity scores serve as emissions in an HMM, and Viterbi decoding with order-aware transition constraints selects a step-aligned keyframe sequence that preserves the temporal progression of the activity. The proposed method retains only 1.72% of the input frames while achieving a Step F1 score of 85.63%. On the question-answer benchmark, the highlights reduce input frames by 31.9% and total token usage by 24.3%, with a $\Delta = 0.167$ decrease in answer quality. These results demonstrate the effectiveness of goal-oriented highlight extraction as a compact episodic memory representation for contextual AI, supporting long-horizon reasoning for agentic AI under limited computational and contextual budgets.

CCS Concepts: • **Human-centered computing** → **Ubiquitous and mobile computing**.

Additional Key Words and Phrases: Egocentric Video, Wearable, Contextual AI, Video Summary, Video Highlight Extraction, AI agent

Authors' Contact Information: [Kaylee Yaxuan Li](#) , University of Michigan, Ann Arbor, Michigan, USA and Reality Labs Research, Meta, Redmond, Washington, USA, yaxuanli@umich.edu; [Ke Sun](#), University of Michigan, Ann Arbor, Michigan, USA, kesuniot@umich.edu; [Kang G. Shin](#), University of Michigan, Ann Arbor, Michigan, USA, kgshin@umich.edu; [Alanson P. Sample](#), University of Michigan, Ann Arbor, Michigan, USA, apsample@umich.edu; [Amy Karlson](#), Reality Labs Research, Meta, Redmond, Washington, USA, akkarlson@gmail.com; [Eric Whitmire](#), Reality Labs Research, Meta, Redmond, Washington, USA, ewhitmire@meta.com; [Yifan Wang](#), Reality Labs Research, Meta, Redmond, Washington, USA, yifan1@meta.com.

Please use nonacm option or ACM Engage class to enable CC licenses



This work is licensed under a [Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License](#).

© 2026 Copyright held by the owner/author(s).

ACM 2474-9567/2026/12-ART230

<https://doi.org/10.1145/3857925>

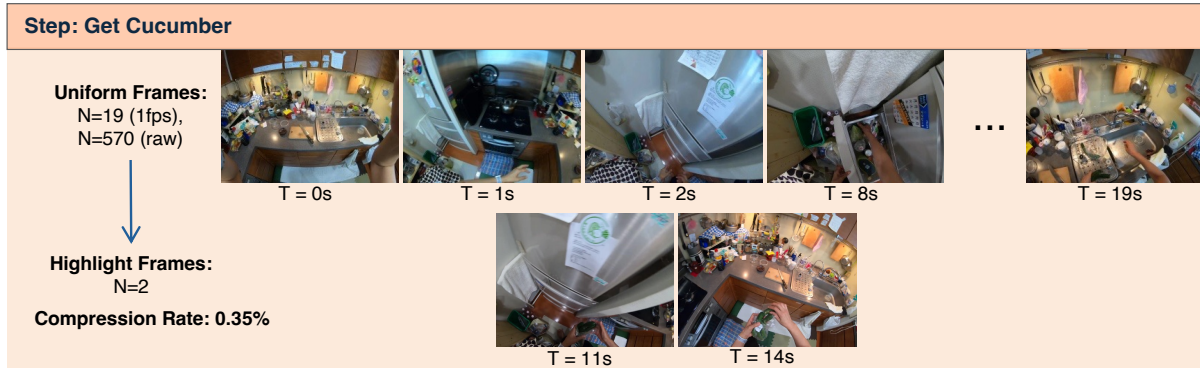


Fig. 1. Comparison of uniform frame sampling and goal-oriented highlight selection. In this example, the proposed method retains two frames that capture the goal-relevant events from a 570-frame input sequence, yielding a retention rate of 0.35%, whereas uniform sampling preserves many redundant observations.

ACM Reference Format:

Kaylee Yaxuan Li, Ke Sun, Kang G. Shin, Alanson P. Sample, Amy Karlson, Eric Whitmire, and Yifan Wang. 2026. GOHi: Goal-Oriented Highlight Extraction for Long-Form Egocentric Video. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 10, 4, Article 230 (December 2026), 27 pages. <https://doi.org/10.1145/3857925>

1 INTRODUCTION

Wearable AI assistants can augment human capabilities by acting as an external memory and providing context-aware guidance grounded in daily life [2, 5, 79, 90], with applications ranging from everyday task completion to support for people experiencing memory loss. First-person video captured by wearable cameras offers a natural source of context for such assistance by preserving visual evidence of what the user has seen and done during everyday activities [4, 17, 51, 61, 79]. However, raw egocentric recordings are continuous and uncured, with meaningful actions sparsely distributed amid static scenes, transitional footage, repeated observations, and incidental interactions [13]. As recording duration increases, the volume of visual data grows while relevant evidence remains sparsely distributed [39]. This mismatch between accumulating video and the limited computational and context budgets of downstream models makes footage difficult to use directly as a long-term memory representation: redundant frames increase storage and inference costs, consume model context, and obscure the evidence needed for retrieval and reasoning. Constructing episodic memory therefore requires a compact representation that preserves essential semantic and temporal evidence while discarding irrelevant content.

Representative keyframes and highlight sequences offer a practical means of constructing such a compact representation. Conventional video summarization methods optimize visual interestingness, representativeness, diversity, or coverage [11, 25, 26]. These criteria, however, do not establish whether a selected moment is semantically important to the user’s activity. Object-centric and action-centric approaches incorporate higher-level semantic cues [34, 53], but identifying salient entities or individual actions does not necessarily capture how those events contribute to a broader activity. Query-focused summarization and moment retrieval condition frame or segment selection on user-supplied text [35, 50, 64], but query formulation does not explicitly represent the ordered sequence of actions through which the overall experience unfolds.

Many everyday activities naturally exhibit goal-directed and hierarchical structure: a broader objective is accomplished through an ordered sequence of steps and lower-level actions [67]. This structure provides the semantic and temporal context needed to determine which observations should be retained. In activities such as

preparing a meal, commuting to work, or locating an object, an observation is important not simply because it is visually distinctive, but because it provides evidence of progress toward the intended outcome. For contextual AI, recognizing *what* occurred is therefore insufficient without also capturing *why* each action matters to the intended goal and *how* the activity progresses over time [42, 53]. Ordered goal steps provide a natural abstraction for selecting and organizing this evidence. For example, a small set of frames showing, e.g., cutting vegetables, heating a pan, and frying the ingredients may preserve the essential progression of meal preparation. As illustrated in Figure 1, two goal-relevant frames capture the essential events in a 570-frame video, whereas uniform sampling retains several redundant observations.

Accordingly, we propose GOHi, a goal-oriented framework that transforms continuous egocentric video into a compact, temporally coherent sequence of representative keyframes. Rather than treating the user’s goal as a single retrieval query, the framework represents it as an ordered sequence of steps and organizes the retained visual evidence around its progression. This formulation distinguishes goal-relevant evidence from incidental and redundant content while representing an extended activity through a small set of semantically meaningful observations. The framework consists of two stages. First, an open-vocabulary vision–language encoder computes semantic compatibility between individual video frames and natural-language descriptions of the goal steps, identifying candidate keyframes for each step. Second, a global temporal optimization module formulates frame–step alignment as a hidden Markov model and applies Viterbi decoding to select the highest-scoring keyframe sequence. Order-aware transition constraints require the selected frames to follow the order of the goal steps, thereby resolving locally plausible but globally inconsistent matches. Together, semantic matching and structured temporal decoding produce a compact highlight sequence that preserves both goal-relevant visual evidence and the temporal progression of the activity.

The evaluation is conducted on 708 videos from the Ego4D Goal-Step benchmark, spanning 302.1 hours. For a comprehensive assessment, grounding accuracy is reported across the range of compression rates obtained by sweeping the candidate-retention percentile $P \in \{2, 4, 6, 8, 10, 15\}\%$, yielding realized compression rates of 0.87%–2.22%. The proposed framework supports two modes: Intention-Primed Alignment combines visual evidence with coarse Gaussian prior guidance, whereas Visual Self-Alignment derives the alignment from visual–semantic evidence. At ($P=8\%$), Intention-Primed Alignment retains only 1.72% of the original video, corresponding to a 58 $\times$ reduction in frames, while achieving 85.63% Step F1 and 62.36% Frame Precision. The evaluation further includes ablation studies of key components, comparisons with state-of-the-art methods, and an analysis of runtime and GPU efficiency. Downstream evaluation on MultiHop-EgoQA question-answer benchmark using ChatGPT-5.5 further shows that the extracted highlights reduce frames input by 31.9% and total token usage by 24.3%, with only a $\Delta = 0.167$ decrease in answer quality on a five-point scale. Together, these results show that goal-oriented highlights remove substantial visual redundancy while preserving the evidence needed for downstream contextual reasoning. Accordingly, this work presents a goal-oriented framework that transforms continuous egocentric video into a compact, temporally coherent sequence of representative keyframes. The main contributions are summarized as follows:

- A formulation of goal-oriented highlight extraction for constructing compact episodic memory from long-form egocentric video.
- A training-free framework that combines goal-step-guided semantic frame selection using a vision–language encoder with HMM-based global temporal optimization via Viterbi decoding.
- A comprehensive evaluation on 708 videos across six candidate-retention levels shows that the method can retain only 1.72% of the original frames while achieving an 85.63% Step F1 score.
- Downstream question-answering evaluation demonstrates that the extracted highlights reduce visual input and token usage while maintaining comparable answer quality.

2 RELATED WORK

2.1 Egocentric Video Understanding and Contextual AI

Egocentric video provides a dominant source for everyday human activity, supporting applications in augmented reality, assistive computing, health monitoring, and robot learning. Unlike third-person video, it captures activity from the camera wearer’s perspective and preserve contextual cues such as object affordances, hand–object interactions, environmental and spatial layout, and the temporal progression of daily tasks. Early datasets, including UT Egocentric [34], ADL [53], GTEA [20], and first-person social-interaction datasets [19], established foundational tasks in activity recognition, object discovery, and social behavior understanding. However, their limited duration, environmental diversity, and procedural complexity constrained the study of long-horizon activity. Subsequent benchmarks expanded both the scale and scope of egocentric understanding. EPIC-KITCHENS [10] advanced fine-grained modeling of procedural actions in cooking; EgoHands [3] and MECCANO [57] emphasized hand–object interactions; and Assembly101 [63] supported multi-view understanding of procedural activities. Large-scale datasets such as Ego4D [22] and Ego-Exo4D [23] further broadened the field toward long-horizon activity understanding, episodic memory, and cross-view reasoning. EgoLife [79] introduced week-long, multimodal recordings with long-context question-answering tasks designed around everyday assistance.

Recent benchmarks examine long-video understanding and memory-dependent embodied reasoning [45, 46, 74, 78]. In parallel, contextual AI systems construct persistent memory from extended egocentric observations to support recall, planning, and personalized assistance. Embodied VideoAgent [17] builds dynamic scene memory from egocentric video and embodied sensors. Other systems increasingly combine large multimodal models with retrieval-augmented generation (RAG) to support clip-level understanding, long-horizon evidence retrieval, and question answering [36, 59, 79]. These systems retrieve clips, scenes, or structured memory records for downstream queries but do not explicitly use the structure of achieving a goal step by step to guide both evidence selection and temporal organization. Consequently, relevant moments may be retrieved without capturing an activity’s stepwise progression. Goal-oriented highlight extraction addresses this complementary objective with ordered goal steps to select explicit frames and organize them into a compact, temporally coherent visual memory.

2.2 Semantic Video Summarization: From Visual Features to Visual Language Models

Video summarization condenses long-form video into compact representations that preserve important content while reducing redundancy. Early methods formulated summarization as unsupervised subset selection, using low-level cues such as color, motion, and edge statistics to identify salient frames or segments [25, 32, 34, 54, 69]. These approaches optimized relevance, representativeness, and diversity, often through submodular objectives that balanced coverage and redundancy [26]. However, visual saliency may not reflect semantic importance, particularly for meaningful but visually subtle or static activities. Subsequent work learned frame importance directly from data. Recurrent models, including LSTMs, captured temporal dependencies and structural consistency between source videos and their summaries [44, 86], while attention-based architectures enabled more flexible temporal modeling [16]. Despite these advances, importance remained largely inferred from visual patterns rather than explicitly grounded in the semantic content of the video. A complementary line of work used human responses to estimate importance. In sports and online video, frequently replayed segments, peaks in time-aligned live-chat activity, and concentrations of viewer comments have been used to identify moments that attract collective attention [28, 31, 68, 80]. Individual-level behavioral and physiological signals, including gaze [77], facial expression, head motion, and EEG [52], have similarly served as proxies for perceived relevance. However, acquiring these signals requires additional sensing and reliable continuous measurement, limiting their practicality for long-term video summarization.

Large-scale vision–language encoders have brought explicit semantic grounding to video summarization [41, 55, 58, 66, 81, 87]. Models such as CLIP [56], ALIGN [30], and SigLIP [71, 84] jointly embed visual content and

language, enabling open-vocabulary frame–text matching for video–text retrieval [18, 43, 73] and language-based temporal localization [21, 29]. Query-focused video summarization conditions frame selection on user-specified text, beginning with egocentric video [64, 65] and extending to attention-based models [76], language-guided summarization [50], and query-conditioned moment and highlight detection [35]. Interactive methods further model user intent as an adjustable distribution over query concepts [75]. However, most approaches treat summarization as retrieval or ranking, using a single query to identify individually relevant moments. This formulation is insufficient for extended goal-directed activities, where summary quality depends not only on whether each moment is relevant, but also on how selected events relate and progress over time. The problem is particularly pronounced in egocentric video, where repeated actions and visually similar scenes can produce many locally relevant but procedurally distinct moments. Representing intention as an ordered sequence of goal-steps provides this missing task structure, guiding both the selection and temporal organization of evidence into a coherent procedural summary.

2.3 Weakly Supervised Temporal Alignment with Structured Decoding

Representing user intention as an ordered sequence of goal steps makes highlight extraction closely related to aligning those steps with temporally ordered visual evidence. Sequence alignment is a classical problem, with foundational formulations including dynamic time warping [62] and connectionist temporal classification [24]. In video understanding, this primitive underlies weakly supervised action alignment, where an ordered transcript of actions is given but frame-level boundaries are not: HMM-based methods align frames to transcript states via Viterbi decoding [33], later combined with neural classifiers trained through the alignment [60] and with differentiable alignment objectives [7, 38]. A parallel line applies these ideas to instructional video, where the transcript consists of natural-language step descriptions [15, 27, 40, 48, 89], most recently without training via large multimodal models [82]. Notably, Drop-DTW [14] extends monotone alignment with an explicit skip mechanism, accommodating real executions in which some steps are absent or interrupted. These studies establish ordered temporal alignment as a mechanism for locating actions or procedural steps in video. This work applies the same general paradigm to a different objective: training-free, intention-conditioned compression of egocentric video. Similarities from a pretrained vision–language encoder define the frame–step emission scores, while Viterbi decoding determines which frames are retained. Rather than returning frame labels or temporal boundaries as the final output, the decoded alignment directly constructs a compact highlight set containing temporally ordered evidence for the user’s goal.

3 SYSTEM DESIGN

This work introduces an intention-driven framework that distills raw footage into a concise, narrative-preserving highlight sequence, providing a high-level representation of long, unstructured egocentric video. In contrast to methods based on low-level saliency or isolated object retrieval, the proposed approach grounds frame selection in the user’s goal-steps, ensuring that the resulting summary reflects the underlying intent of the activity rather than incidental visual cues. Within this pipeline, the input video is represented as an ordered sequence of frames $V = \{v_1, v_2, \dots, v_N\}$; Because ordered goal steps encode the stepwise procedure for completing a task, user intention is represented as an ordered sequence of textual steps, denoted as $T = \{t_1, t_2, \dots, t_M\}$. The objective is to select a small subset of frames that provides essential visual evidence that reassembles each of these steps while promoting consistency with their temporal progression.

Figure 2 presents the overall architecture of the pipeline. First, **cross-modal embedding** maps sampled video frames and ordered textual goal steps into a shared representation space for direct frame–step comparison (Section 3.1). Second, **semantic matching and filtering** computes frame–step similarities and retains strongly aligned frames as candidates (Section 3.2). Finally, **temporal alignment and decoding** combines semantic

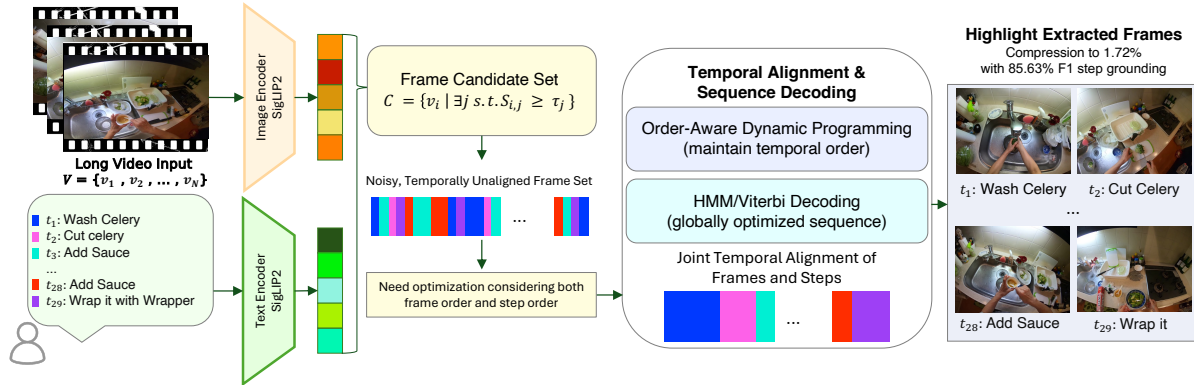


Fig. 2. System overview of the proposed goal-oriented highlight extraction framework. Long egocentric video frames and user’s textual goal-steps are embedded into a shared semantic space using SigLIP 2. Frames with high semantic alignment to each step form a candidate keyframe set. An order-aware Viterbi decoder then encourages consistency between frame order and step order, producing a compact, temporally coherent highlight sequence.

similarity with mode-specific temporal guidance and order-aware transitions to assign candidate frames to the ordered steps (Section 3.3). Global sequence optimization helps resolve locally ambiguous and repeated actions, producing a temporally coherent highlight sequence.

3.1 Vision-Language Embedding

Given sampled video frames and textual goal steps, this stage maps both modalities into a shared representation space. The resulting embeddings enable semantic frame–step matching and provide the inputs to candidate selection. To bridge the semantic gap between low-level visual signals and high-level user steps, this work employs SigLIP 2, a state-of-the-art multilingual vision–language encoder [71]. Unlike earlier models such as CLIP [56], which rely on a softmax-based contrastive objective, SigLIP 2 adopts a sigmoid loss that treats image–text matching as a collection of independent binary classification problems [84]. This design supports scalable training and stronger semantic alignment, which is particularly important for resolving fine-grained distinctions in everyday activities (e.g., distinguishing “washing celery” from “holding celery”). Beyond global image–text alignment, SigLIP 2 adopts a unified training recipe that yields both global and dense spatial representations. This pipeline uses the So400M (NaFlex) variant, whose three architectural components underpin the quality of the extracted embeddings. These components collectively address the visual difficulties of egocentric video, including its wide field of view, small objects in cluttered scenes, and heavy hand occlusion, as detailed below.

- (1) **NaFlex Vision Encoder.** The vision encoder converts input frames into semantic embeddings. Standard Vision Transformers (ViTs) resize images to a fixed square resolution (e.g., 224×224), which can introduce aspect-ratio distortion or cropping that removes peripheral visual context. SigLIP 2 addresses this limitation through the NaFlex architecture, which combines variable sequence-length support (FlexiViT) [6] with aspect-ratio-aware preprocessing (NaViT) [12]. Images are resized to a non-square patch grid that aligns with the ViT tokenizer while keeping distortion minimal and the token length bounded. The learned two-dimensional positional embeddings are bilinearly interpolated to match this grid, and padding tokens are masked during attention. This design preserves the wide field of view characteristic of egocentric imagery while enabling stable and efficient encoding.

- (2) **Multilingual Text Encoder and MAP Pooling.** To encode user steps, SigLIP 2 employs a Transformer-based text encoder that differs from CLIP-style encoders in two key aspects. First, it adopts the multilingual Gemma tokenizer with a 256k vocabulary, enabling robust handling of diverse, open-vocabulary step descriptions across languages [70]. Input text is lower-cased and tokenized with a maximum sequence length of 64 tokens. Second, instead of aggregating token features with a fixed [CLS] token or simple mean pooling, the encoder applies a Multi-Head Attention Pooling (MAP) head [83]: a learned attention mechanism that assigns data-dependent weights to token-level representations before projecting the sequence into the joint vision–language embedding space. The resulting representation emphasizes semantically informative terms, such as action verbs and objects, while down-weighting less informative tokens.
- (3) **Unified Training Recipe for Dense Understanding.** Beyond architectural design, SigLIP 2 derives its performance from a unified training recipe that extends the core sigmoid image–text objective with auxiliary supervision to preserve dense and spatially grounded representations. Standard contrastive models often emphasize global alignment [56], producing representations that can underrepresent fine-grained spatial structure critical for reasoning about detailed interactions [49, 72]. SigLIP 2 addresses this limitation through two complementary mechanisms.
- **Location-Aware Captioning (LocCa).** During pretraining, the model is jointly optimized with a Transformer-based decoder for grounded captioning and referring-expression tasks. This objective requires the image encoder to retain spatially localized features that support region-level grounding, rather than collapsing visual information into a single global embedding. As a result, the learned representations better preserve spatial structure across the image.
 - **Self-Distillation with Masked Prediction.** Following the SILC [49] and TIPS [47] protocols, SigLIP 2 incorporates a self-distillation stage in which a student network predicts the representations of a teacher network from partially masked image patches. This local-to-global correspondence learning yields fine-grained representations of image regions, which is particularly important for recognizing small objects and inferring occluded content in visually cluttered egocentric scenes with heavy hand occlusion.

Once the SigLIP 2 architecture is established, the next step is to encode each video frame and the user-provided step text into a shared embedding space using SigLIP 2 for inference. During inference, each video frame v_i is processed by the NaFlex Vision Encoder ($\mathcal{E}_{\text{img}}$), while each textual step t_j is processed by the Text Encoder ($\mathcal{E}_{\text{txt}}$). Both modalities are projected into a shared d -dimensional embedding space:

$$\mathbf{v}_i = \mathcal{E}_{\text{img}}(v_i), \quad \mathbf{t}_j = \mathcal{E}_{\text{txt}}(t_j), \quad (1)$$

where $\mathbf{v}_i, \mathbf{t}_j \in \mathbb{R}^d$ denote the visual and textual embeddings, respectively. These embeddings serve as input to the subsequent stage.

3.2 Frame Candidate Selection

Given the visual and textual embeddings, the next stage retains a subset of frames that are semantically relevant to the user’s goal-steps, through semantic alignment computation followed by percentile-based filtering.

(1) **Semantic Alignment Calculation.**

The semantic correspondence between video frames and textual steps is represented by a pairwise similarity matrix $\mathbf{S} \in \mathbb{R}^{N \times M}$. Each entry is computed as the cosine similarity between the corresponding visual and textual embeddings:

$$S_{i,j} = \frac{\mathbf{v}_i^\top \mathbf{t}_j}{\|\mathbf{v}_i\|_2 \|\mathbf{t}_j\|_2}. \quad (2)$$

A higher value of $S_{i,j}$ indicates stronger semantic correspondence between frame v_i and step t_j .

(2) **Step-Wise Candidate Frame Filtering.**

To discard redundant content such as camera motion, transitions, and activities with weak alignment to the given steps, a *candidate frame set* is constructed by adaptive per-step filtering. For each step t_j , a dynamic threshold τ_j is set at the $(100 - P)$ -th percentile of that step's similarity scores, so that the top $P\%$ of frames are retained:

$$\tau_j = Q_{1-P/100} \left(\{S_{i,j}\}_{i=1}^N \right), \quad (3)$$

where $Q_q(\cdot)$ denotes the q -th quantile and P the percentage of frames retained per step. The candidate set for step t_j collects the frames that pass this threshold:

$$C_j^P = \{v_i \mid S_{i,j} \geq \tau_j\}. \quad (4)$$

To guarantee a minimum candidate pool for each step, this set is augmented with the $K_{\min} = 8$ highest-scoring frames, denoted by C_j^K : $C_j = C_j^P \cup C_j^K$. The final candidate set is the union across all steps:

$$C = \bigcup_{j=1}^M C_j. \quad (5)$$

Because filtering is applied independently to each step, the resulting candidate count depends on the video length, number of steps, score distribution, and overlap among the step-specific candidate sets. The retained candidates are ordered by timestamp and indexed as $c_1, \dots, c_{N_c}$, where $N_c = |C|$. This stage reduces the search space for subsequent temporal alignment and sequence decoding.

3.3 Temporal Optimization

Candidate selection identifies semantically relevant frames but does not determine which step each frame represents or whether the assignments follow the activity's progression. Empirical analysis of directly concatenating these candidate frames reveals three reliability issues:

- **Visual Ambiguity and Unclear Decision Boundaries:**

Similarity scores for semantically related actions (e.g., “add soy sauce” versus “add vinegar”) cluster closely, as such scenes share near-identical visual elements: a bottle, a kitchen counter, and a pouring motion. Scores for competing steps often differ by margins below 0.01, leaving little separation to exploit. In this low-margin regime, greedy per-frame selection amplifies noise, assigning frames to steps on the basis of minor fluctuations in the embedding space rather than semantic correspondence.

- **Overshadowing of Short Actions:**

Percentile-based selection allocates candidates in proportion to video length rather than step duration. A brief step such as “peeling” may last only a few seconds, yet a fixed threshold (e.g., top 10%) retains a far larger set of frames for it. Because similarity rarely drops sharply at action boundaries, the remaining slots are filled with visually similar but uninformative frames, such as static holding or transitions. These frames outnumber the valid ones, making short actions difficult to isolate in the subsequent alignment stage.

- **Ambiguity in Repetitive Actions:**

Independent matching cannot disambiguate recurring steps. When an action such as “cutting” appears several times in the step sequence, a single frame carries no information about *which* occurrence it belongs to. The result is temporal misalignment: a frame from a later cutting stage is assigned to an earlier occurrence of the step, violating the order of the activity.

To address these issues, frame selection is reformulated as a global alignment problem. The difficulty stems from the *many-to-many correspondence* between visual signals and semantic steps: a single step such as “wash celery” spans many consecutive frames, while an individual frame may align with several distinct steps, particularly when the same action recurs at different stages of the activity. Independent per-frame retrieval therefore cannot resolve the assignment; it requires a globally optimal path that maps frames to steps while promoting consistency

with their temporal order. This is operationalized by modeling the video as a Hidden Markov Model (HMM) and decoding it with dynamic programming. Sequence alignment between an ordered transcript and video is well established in weakly supervised action segmentation [33, 60]. This work adopts the same decoding paradigm with a forward-biased transition structure. The formulation separates the latent step sequence from noisy visual observations:

- **Latent State Representation.** The state space follows the ordered sequence of step occurrences $T = \{t_1, \dots, t_M\}$, together with a background state b that absorbs frames belonging to no step. An activity that recurs appears as several entries, each with its own state, so repetition is represented by construction rather than collapsed into a single category. At each time step k , the decoder occupies a latent state $q_k \in \{1, \dots, M\} \cup \{b\}$, indicating the active step, and the frame embedding $\mathbf{v}_k$ serves as the observation. Decoding begins in the first step state or in the background state; all other states are disallowed at $k = 1$. Formally, the initial-state scores are $\pi(1) = 0.5$, $\pi(b) = 0$, and $\pi(j) = -\infty$ for $j \in \{2, \dots, M\}$. The emission score for assigning frame k to step j is derived from the semantic alignment score $S_{k,j}$ through the calibration described below. This formulation is robust to local visual noise: the state at time k is determined not by the instantaneous match $S_{k,j}$ alone, but by the total score of the alignment over the complete candidate sequence.
- **Emission Calibration and Temporal Guidance.** Similarity ranges vary across textual steps, so candidate-level scores are standardized independently for each step, yielding $Z_{k,j}$. Specifically, $Z_{k,j} = (S_{k,j} - \mu_j) / (\sigma_j + \epsilon)$, where μ_j and σ_j are computed over the candidate scores for step t_j . A soft temporal guidance term $G_{k,j}$ then adjusts the visual evidence according to the expected occurrence of step t_j . The combined scores are converted into step-state emission scores using a temperature-scaled log-softmax:

$$\psi_k(j) = \log \text{softmax}_j \left(\frac{Z_{k,j} + \lambda G_{k,j}}{\tau} \right), \quad (6)$$

where λ is a mode-specific temporal-guidance weight. The background state is assigned a score below the least-supported step at each candidate, allowing weakly aligned frames to remain unassigned. Within the shared decoding framework, the two modes differ in how temporal guidance is obtained. In Intention-Primed Alignment, $G_{k,j}$ is a bounded Gaussian centered at a coarse temporal cue a_j associated with step t_j , such as the user's approximate recollection of when the step occurred:

$$G_{k,j} = \exp \left(-\frac{(t(c_k) - a_j)^2}{2\sigma_a^2} \right), \quad (7)$$

where $t(c_k)$ denotes the timestamp of candidate c_k , and σ_a controls the temporal bandwidth of the cue. The Gaussian regularizes the emission scores rather than determining the alignment independently. Final frame-step assignments are jointly determined by visual-language similarity and global sequence optimization under the order-aware Viterbi decoder.

In Visual Self-Alignment, the Gaussian is replaced by a self-estimated temporal prior: ordered step regions are inferred from visual-language scores over the sampled timeline, and the resulting prior regularizes the emission scores alongside visual evidence under the same decoding framework.

- **Order-Aware Transition Structure.** Temporal progression is modeled through the transition scores $\phi(\cdot, \cdot)$ between states. A step state may remain active ($\phi = +0.50$), modeling variable action duration, or advance to its immediate successor ($\phi = 0.00$). Forward skips to a later step carry a fixed penalty ($\phi = -2.00$), allowing the decoder to pass over steps that lack visual support rather than forcing spurious assignments, following the motivation of skip-tolerant alignment [14]. Direct backward transitions between step states are prohibited ($\phi = -\infty$). Transitions into the background state score -0.50 , transitions from it to any step

state +0.20, and remaining in it +0.50. The background state therefore serves two functions: it absorbs frames that match no step, and it provides the only path by which the decoder can return to an earlier step. The transition structure is therefore forward-biased between consecutive active assignments, while transitions through the background state may return to an earlier step.

- **Global Optimization via Viterbi Decoding.** The number of possible state sequences grows exponentially with the length of the candidate sequence, making exhaustive search infeasible. Because the objective decomposes into per-frame emission terms and pairwise transition terms between consecutive states, it admits an exact dynamic-programming solution. Viterbi decoding computes the globally highest-scoring state sequence $y_{1:N_c}^*$ in $O(N_c M^2)$ time, determining all frame-step assignments jointly rather than independently, under the order-aware transition structure.
- **Optimization Objective.** Let $\mathcal{Y} = \{1, \dots, M, b\}$ denote the M step states and the background state b . The optimal alignment is the state sequence that maximizes the total sequence score:

$$y_{1:N_c}^* = \arg \max_{y_{1:N_c} \in \mathcal{Y}^{N_c}} \left[\pi(y_1) + \sum_{k=1}^{N_c} \psi_k(y_k) + \sum_{k=2}^{N_c} \phi(y_{k-1}, y_k) \right], \quad (8)$$

where $\pi(y_1)$ is the initial-state score, $\psi_k(y_k)$ is the emission score for assigning candidate frame c_k to state y_k , and $\phi(y_{k-1}, y_k)$ is the transition score between consecutive states. The two modes share this decoding objective and transition model but construct the emission scores differently. Intention-Primed Alignment uses coarse temporal cues, whereas Visual Self-Alignment first estimates ordered temporal regions from the full visual sequence and uses them to derive visual temporal guidance. The terms constitute an additive sequence score rather than a normalized joint probability; accordingly, Viterbi decoding maximizes the total alignment score rather than a likelihood. This global objective evaluates each assignment in the context of the complete candidate sequence, reducing errors caused by locally ambiguous frame-step similarities.

- **Step Coverage and Missing Steps.** Forward skips allow the decoder to bypass steps with insufficient visual support. After decoding, any step with no assigned candidate is given the highest-scoring frame from its step-specific candidate set. These fallback selections are marked separately from Viterbi assignments. Thus, complete step coverage is enforced during post-processing rather than by the transition structure.

Together, these stages transform the sampled video into a compact sequence of step-linked highlights.

4 EVALUATION PROTOCOL

This section specifies the evaluation protocol: Section 4.1 describes the evaluation population, Section 4.2 defines the metrics, Section 4.3 presents the implementation details and hyperparameters, and Section 4.4 introduces the compared methods and budget-matching procedure used for the budget-constrained baselines.

4.1 Dataset

The GOHi framework is evaluated on the Ego4D Goal-Step benchmark [67], which targets hierarchical procedural understanding in long-form egocentric video. The wearer cooks naturally, without following a script. Visual evidence of the ordered procedure is therefore interspersed with redundant footage, transitions, and off-task activity, reflecting realistic operating conditions for a wearable assistant. Goal-Step organizes activities into three levels: high-level goals (e.g., “Make an omelet”), intermediate steps (e.g., “Crack eggs”), and fine-grained substeps (e.g., “Pick up egg”). The annotations provide ordered, temporally localized procedural events distributed throughout long recordings. This structure provides a suitable testbed for intention-driven summarization, which must identify sparse visual evidence for an ordered procedure amid substantial redundancy, transitions, and activity unrelated to the evaluated steps.

The dataset split provides 583 training and 134 validation videos. Because the method is training-free, the training split poses no contamination risk and both splits are evaluated. Nine videos yield no valid step occurrences under the evaluation filtering protocol and are excluded from every condition, leaving 708 videos (131 validation, 577 training) totaling 302.1 hours. The evaluated videos are long and procedurally diverse, with a median duration of 19.3 minutes (IQR: 10.9–31.6) and a median of 27 annotated substeps per video (IQR: 12–49). Across 27,521 evaluated substeps, the median duration is 11.6 seconds (IQR: 5.3–25.0). The duration spans 1.6 seconds at the 5th percentile to 46.8 minutes at the maximum; 23.4% last fewer than five seconds. These statistics reflect substantial variation in both procedural complexity and temporal scale across the benchmark.

Procedural activity in continuous egocentric capture is rarely a linear pass through a plan. People revisit steps, interleave them with unrelated activity, and abandon and resume tasks. This population reflects that. Of the 708 evaluated videos, 584 (82.5%) contain at least one repeated step, and 10,066 of the 27,521 evaluated substeps (36.6%) revisit a step appearing earlier in the same video; the figure is insensitive to the matching rule, as lemmatized matching yields 37.6%. The input construction preserves these repeated occurrences. The state sequence is defined over ordered step occurrences rather than unique activity categories. A recurring activity contributes a separate state for each occurrence, allowing repetitions to be represented without collapsing them into a single category. Preserving occurrence order therefore places no restriction on how often an activity recurs. Off-task activity is likewise present rather than filtered away. Step lists exclude optional and irrelevant entries, together with five categories that occur incidentally across goals rather than as part of any procedure (*use phone*, *use computer*, *write a note*, *set the dining table*, *do laundry*). The exclusion applies to the step list alone; the corresponding video intervals are not removed from the visual input. Activity from excluded categories therefore persists as a distractor during frame selection rather than being removed from the input. Annotated off-task segments appear in 50 videos, accounting for 62.5% of runtime in one example video.

4.2 Evaluation Metrics

Evaluating procedural highlight frames extraction requires balancing three objectives: compactness, procedural coverage, and temporal precision. Accordingly, four metrics are reported: Compression Rate, the fraction of the original recording retained after selection; Step Recall, the proportion of stated steps for which the selected frames provide evidence; Step F1-Score, the balance between that coverage and steps predicted but not located; and Frame Precision, the proportion of retained frames falling within their assigned intervals. Together they characterize the trade-off between compactness, procedural coverage, and temporal alignment.

- **Compression Rate (Compactness).** Compression Rate measures summary compactness and is defined as the ratio of selected highlight frames to the total number of frames in the original 30 FPS recording:

$$\text{Compression Rate} = \frac{\sum_v N_{\text{selected}}^{(v)}}{\sum_v N_{\text{original}}^{(v)}}, \quad (9)$$

where $N_{\text{original}}^{(v)}$ is the raw frame count of video v before temporal downsampling. The ratio is aggregated at the dataset level, as the sum of selected frames over the sum of raw frames, so that videos contribute in proportion to their length rather than equally. Lower values indicate stronger temporal compression.

- **Step Recall (Coverage).** Step Recall measures procedural coverage. A ground-truth step occurrence s_i is considered recovered if at least one selected frame assigned to s_i has a timestamp within its annotated temporal interval. Step Recall is defined as

$$\text{Step Recall} = \frac{1}{N} \sum_{i=1}^N \#[\text{recovered}(s_i)], \quad (10)$$

where N is the number of evaluated step occurrences in the input plan.

- **Step F1-Score.** Step F1-Score summarizes step-level coverage and correctness as the harmonic mean of Step Recall and Step Precision. A step occurrence is considered predicted if at least one selected frame is assigned to it. Step Precision is the fraction of predicted step occurrences for which at least one assigned frame lies within the corresponding ground-truth interval:

$$\text{Step Precision} = \frac{\# \{s_i : s_i \text{ is predicted and correctly recovered}\}}{\# \{s_i : s_i \text{ is predicted}\}}. \quad (11)$$

The Step F1-Score is then

$$\text{Step F1} = \frac{2 \cdot \text{Step Precision} \cdot \text{Step Recall}}{\text{Step Precision} + \text{Step Recall}}. \quad (12)$$

- **Frame Precision (Frame-Level Alignment).** Frame Precision measures temporal alignment at the frame level. A selected frame assigned to step occurrence s_i is considered correct if its timestamp lies within the annotated interval of s_i . It is defined as

$$\text{Frame Precision} = \frac{N_{\text{correct}}}{N_{\text{selected}}}, \quad (13)$$

where N_{correct} is the number of selected frames that fall within the annotated interval of their assigned step occurrence. Matching uses zero temporal slack and no global temporal offset: a selected frame contributes only if its timestamp lies inside the annotated interval.

Compression Rate is aggregated over the full dataset as defined above. Step Recall, Step F1-Score, and Frame Precision are computed separately for each video and then averaged over the 708 evaluated videos. Matching uses zero temporal slack and no global temporal offset: a selected frame contributes only if its timestamp lies inside the annotated interval. A video in which no selected frame falls within any annotated interval yields an undefined Step Precision and Step F1; such videos contribute zero rather than being excluded from the average, so that a method is not rewarded for failing completely rather than partially. Together, these metrics characterize the trade-off between summary compactness, procedural coverage, and temporal alignment. Compression Rate quantifies the reduction in visual input, while the step-level and frame-level metrics measure whether the compressed summary preserves the procedural evidence required for subsequent review and reasoning.

4.3 Experimental Settings

Following common practice in long-form video understanding, videos are subsampled to 1 FPS. The candidate-retention percentage is evaluated at $P \in \{2, 4, 6, 8, 10, 15\}\%$, providing operating points with different compression-coverage trade-offs. For the emission model defined in Section 3.3, the softmax temperature is fixed at $\tau = 0.7$. Intention-Primed Alignment uses a temporal-guidance weight of $\lambda_{\text{IPA}} = 2.0$ and a Gaussian bandwidth of $\sigma_a = 30$ seconds, whereas Visual Self-Alignment uses $\lambda_{\text{VSA}} = 0.5$. The temporal-guidance weights, Gaussian bandwidth, and softmax temperature were selected on a small ten-video development subset and held fixed across all videos and candidate-retention levels unless explicitly varied in an ablation. No per-video or per-operating-point tuning is performed. As described in Section 3, the Gaussian prior uses the user's approximate recollection as a temporal cue. The Gaussian prior provides only soft guidance, while visual evidence and global decoding determine the final alignment. Handling inaccurate recollections is outside the scope of this work and is discussed in Section 7. In the experiments, the temporal center of each annotated step is used as the proxy anchor for constructing Gaussian prior.

4.4 Compared Methods and Budget Matching

Two evaluation regimes are used. All methods implemented in this paper are evaluated on the same 708 videos using identical targets and metrics. The comparison set includes four budget-constrained baselines—Uniform Sampling, GPT-5.5 Semantic Scoring, VSUMM-Style Clustering, and CLIP-Based Independent Retrieval—together with three component ablations and one oracle upper bound. The comparison with Training-Free Online Video Step Grounding [82] follows a separate protocol. Rather than reproducing the published system, the proposed method is evaluated under its reported setting, using the same video set, target construction, and metric. Results from the two protocols are reported separately.

To isolate selection quality from summary length, all methods are evaluated under identical per-video frame budgets. At each operating point, the output of Intention-Primed Alignment defines the per-video budget for the four baselines, preserving the same evaluation population across methods. Visual Self-Alignment is evaluated at the same candidate-retention percentage and therefore has an independently realized output size. Component ablations are evaluated at the same candidate-retention percentage P as Intention-Primed Alignment rather than under a forced output budget.

5 EVALUATION RESULTS

This section evaluates the proposed framework against four complementary keyframe-selection baselines under identical per-video frame budgets. The comparison spans established classical summarization methods and contemporary vision-language approaches: content-agnostic temporal sampling, representative-frame clustering with modern visual features, zero-shot query-conditioned retrieval, and high-capacity multimodal semantic scoring. Together, these baselines position the proposed framework against both canonical and current alternatives for goal-conditioned video highlight extraction. A separate protocol-matched comparison with BaGLM, a recent training-free method for online video step grounding [82], is presented in Section 5.2.

Uniform Sampling isolates the effect of the frame budget: any improvement over it reflect content-aware frame selection rather than from the frame budget itself. It selects frames at fixed temporal intervals without using either visual content or step descriptions. Selected frames are assigned to the ordered steps by partitioning the summary into contiguous, equal-size blocks. This baseline measures the extent to which temporal coverage and frame budget alone recover procedural content.

VSUMM-Style Clustering establishes what visual diversity alone recovers. It represents the classical generic video-summarization paradigm, in which keyframes are selected to capture the visual diversity of a recording. The original VSUMM method clusters hand-crafted visual descriptors [11]; here, k -means is instead applied to SigLIP 2 frame embeddings, and the frames nearest the cluster centroids are retained. This adaptation evaluates the underlying representative-frame principle using modern visual features rather than penalizing the baseline for its original feature representation. Because the selection process does not use the step list, it tests whether generic visual representativeness is sufficient to preserve goal-relevant procedural evidence. Selected frames are assigned to steps using the same block-based rule as Uniform Sampling.

CLIP-Based Independent Retrieval isolates sequence-level structure and exposes the contribution of temporal optimization. It provides a strong zero-shot vision-language baseline for query-conditioned frame selection. Frames are ranked independently for each step using frame-text similarity from CLIP ViT-L/14 at 336 px, and the highest-ranked frames are retained under the shared budget. The baseline reflects the direct retrieval principle underlying CLIP-based query-focused video methods [43, 50], while excluding their learned temporal components. It is the closest non-sequential alternative to the proposed framework: both condition selection on the ordered step list, but independent retrieval does not jointly optimize frame assignments across step occurrences. To separate this difference from encoder choice, Section 5.3 also evaluates independent ranking with SigLIP 2 embeddings.

GPT-5.5 Semantic Scoring provides a high-capacity, general-purpose multimodal baseline. Each sampled frame is evaluated for relevance to the full step list using a rolling context over preceding frames, after which the highest-scoring frames are retained and assigned to their highest-scoring steps. This baseline tests whether general multimodal semantic reasoning is sufficient without an explicit temporal prior, sequence decoder, or step-coverage mechanism.

These baselines isolate four distinct sources of performance: temporal coverage, visual representativeness, independent frame–text retrieval, and general-purpose multimodal reasoning. Their comparison with the proposed framework therefore assesses the additional value of structured, goal-conditioned temporal alignment.

5.1 Results

Table 1 reports the primary comparison at $P = 8\%$, with realized compression rates ranging from 1.57% to 1.77%. Both operating modes of the proposed method outperform all comparison methods. Intention-Primed Alignment provides the strongest performance, with 85.63% Step F1 and 62.36% Frame Precision, while Visual Self-Alignment yields 62.46% Step F1 and 41.98% Frame Precision. The slightly lower compression rates of CLIP-Based Independent Retrieval and VSUMM-Style Clustering arise because their eligible frame sets are insufficient to fill the assigned per-video budget in some videos. The compression rates of the two proposed operating modes are computed independently from their respective retention percentages. The strongest baselines, GPT-5.5 Semantic Scoring and CLIP-Based Independent Retrieval, reach 39.99% and 39.77% Step F1, respectively. These results indicate that organizing visual evidence according to the ordered step sequence provides substantial benefit beyond temporal coverage, generic visual representativeness, and independent semantic matching. Step Coverage and Step F1 coincide for five of the six methods, because these methods assign at least one frame to every input step, making Step Precision equal to Step Recall, and hence Step F1 equal to Step Coverage. CLIP-Based Independent Retrieval leads all baselines in Step Coverage at 39.77%, yet records the lowest Frame Precision at 15.45%, below even content-agnostic Uniform Sampling. Conditioning selection on the step list helps identify which procedural steps are visually present, but does not reliably determine when each step occurs. Frame–text similarity may remain high across multiple visually similar moments, causing independently ranked frames to concentrate on typical appearances of an action rather than the temporal interval of the corresponding step occurrence. This contrast highlights the importance of sequence-level temporal organization beyond independent semantic retrieval.

Moreover, Table 2 and Figure 3 report all six retention levels, spanning compression rates from 0.87% to 2.05% and Step F1 from 71.36% to 86.09% for Intention-Primed Alignment mode extracting highlight frames. Increasing retention produces different frame-level effects across the methods. The two semantic baselines show a clearer precision decline as the budget grows: GPT-5.5 Semantic Scoring falls from 32.86% to 27.75%, while CLIP-Based

Table 1. Highlight extraction performance at the top-8% operating point ($P = 8$). All methods are evaluated on the same 708 videos. Compression is the realized fraction of frames retained from the original 30 FPS stream. Arrows indicate the preferred direction.

Method	Compression (%) ↓	Coverage (%) ↑	Step F1 (%) ↑	Frame Prec. (%) ↑
Intention-Primed Alignment	1.72	85.63	85.63	62.36
Visual Self-Alignment	1.77	62.46	62.46	41.98
GPT-5.5 Semantic Scoring	1.72	34.88	39.99	29.19
CLIP-Based Independent Retrieval	1.57	39.77	39.77	15.45
Uniform Sampling	1.72	33.27	33.27	16.35
VSUMM-Style Clustering	1.63	31.99	31.99	15.95

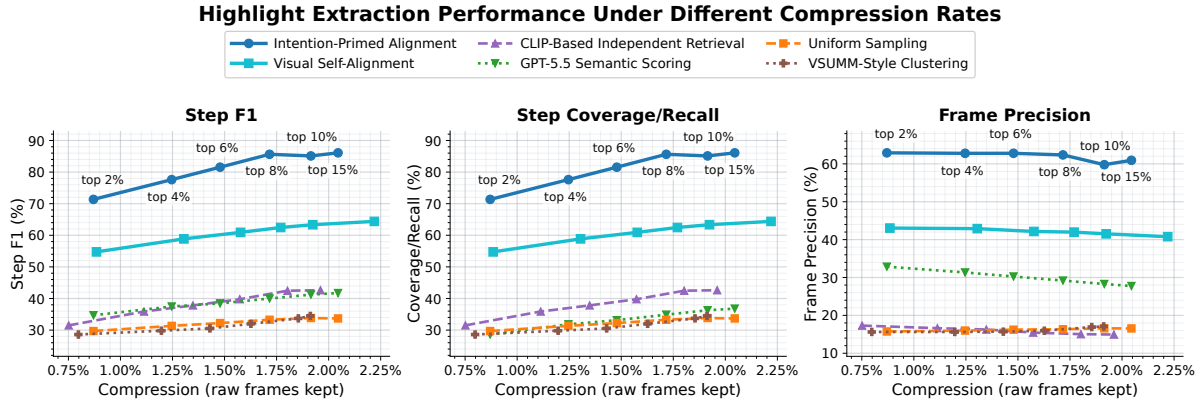


Fig. 3. Highlight extraction performance across varying compression rates on the Ego4D Goal-Step benchmark. The two proposed operating modes are evaluated at different candidate-retention percentages P and compared with GPT-5.5 Semantic Scoring, CLIP-Based Independent Retrieval, Uniform Sampling, and VSUMM-Style Clustering.

Table 2. Highlight extraction performance across all candidate retention levels P on 708 Ego4D Goal-Step videos. Compression Rate is the fraction of the original 30 FPS stream retained.

P (%)	Step Coverage (%) $\uparrow$						Step F1 (%) $\uparrow$						Frame Precision (%) $\uparrow$					
	2	4	6	8	10	15	2	4	6	8	10	15	2	4	6	8	10	15
Compression, IPA (%) $\downarrow$	0.87	1.25	1.48	1.72	1.91	2.05	0.87	1.25	1.48	1.72	1.91	2.05	0.87	1.25	1.48	1.72	1.91	2.05
Compression, VSA (%) $\downarrow$	0.88	1.30	1.58	1.77	1.92	2.22	0.88	1.30	1.58	1.77	1.92	2.22	0.88	1.30	1.58	1.77	1.92	2.22
Intention-Primed Alignment	71.36	77.61	81.55	85.63	85.11	86.09	71.36	77.61	81.55	85.63	85.11	86.09	62.91	62.78	62.79	62.36	59.78	60.95
Visual Self-Alignment	54.75	58.87	60.93	62.46	63.35	64.39	54.75	58.87	60.93	62.46	63.35	64.39	43.05	42.90	42.15	41.98	41.52	40.78
Uniform Sampling	29.70	31.33	32.22	33.27	33.82	33.69	29.70	31.33	32.22	33.27	33.82	33.69	15.74	15.99	16.23	16.35	16.58	16.55
VSUMM-Style Clustering	28.59	29.80	30.52	31.99	33.70	34.48	28.59	29.80	30.52	31.99	33.70	34.48	15.62	15.61	15.70	15.95	16.92	17.08
CLIP-Based Independent Retrieval	31.49	35.88	37.81	39.77	42.46	42.62	31.49	35.88	37.81	39.77	42.46	42.62	17.30	16.62	16.26	15.45	15.04	14.94
GPT-5.5 Semantic Scoring	28.61	31.82	33.12	34.88	36.28	36.79	34.70	37.46	38.46	39.99	41.23	41.72	32.86	31.33	30.26	29.19	28.28	27.75

Independent Retrieval falls from 17.30% to 14.94%. In contrast, Uniform Sampling and VSUMM-Style Clustering remain nearly flat. These trends indicate that adding independently selected semantic matches primarily increases coverage, but also introduces visually plausible frames outside the correct step intervals. Structured alignment maintains substantially higher temporal precision throughout the compression sweep. Overall, these results show that effective highlight extraction requires both a strong visual encoder and structured, goal-conditioned temporal alignment to balance compression, procedural coverage, and frame-level precision.

5.2 Comparison with Training-Free Video Step Grounding

BaGLM [82] is a recent training-free framework for video step grounding that reports comparisons with state-of-the-art specialized grounding methods and strong zero-shot multimodal models. Although video step grounding and highlight extraction both localize natural-language procedural steps in long, untrimmed video, their standard evaluation protocols differ in target construction and metrics. The comparison is therefore reported separately

Table 3. Comparison under the exact Ego4D Goal-Step evaluation protocol used by BaGLM [82]. All methods are evaluated on the same 131-video validation split using the same 2,276 canonical video–category targets and pooled Recall@1 metric. Prior results are reported in the BaGLM study; Visual Self-Alignment is evaluated using its final extracted highlights at $P = 8\%$.

Method	Pooled R@1 (%)
<i>Specialized video step grounding</i>	
VSLNet [85]	24.3
NaSVA [40]	29.1
NaSVA (causal variant) [40, 82]	24.2
<i>Off-the-shelf multimodal models</i>	
LLaVA-OneVision-7B [37]	36.4
Qwen2.5-VL-7B [1]	36.0
InternVL3-8B [88]	44.2
InternVL2.5-8B [9]	45.20
<i>Training-free structured grounding</i>	
BaGLM [82]	43.3
Visual Self-Alignment (ours), $P = 8\%$	58.92

from the main highlight-extraction evaluation. Visual Self-Alignment is used because it most closely matches the BaGLM input setting: both methods operate without task-specific training or externally supplied temporal guidance. Visual Self-Alignment is evaluated using its final extracted highlights at $P = 8\%$.

The protocol is matched in the aspects that determine the evaluation metrics although Visual Self-Alignment operates offline, whereas BaGLM operates online. Both evaluations use the identical 131-video validation split, top-level step descriptions, text canonicalization procedure, and per-video deduplication of canonical step categories. Repeated occurrences of the same category form a single prediction target, while all corresponding annotated intervals remain valid. Each video–category target receives one temporal prediction, which is evaluated using the same interval discretization and pooled Recall@1 computation. The resulting evaluation set contains 2,276 video–category targets, and the locally generated target mapping matches the released BaGLM mapping exactly, with zero mismatches.

As shown in Table 3, Visual Self-Alignment achieves 58.92% pooled Recall@1, outperforming every Ego4D Goal-Step result reported in the BaGLM study. The strongest prior result is obtained by InternVL2.5-8B at 45.20%, yielding an improvement of 13.72 percentage points. Visual Self-Alignment also exceeds BaGLM by 15.62 percentage points and NaSVA, the strongest reported specialized grounding model, by 29.82 percentage points. Notably, the proposed result is computed from the final extracted highlights rather than predictions over the unrestricted video timeline. Under this exactly matched evaluation protocol, Visual Self-Alignment achieves the highest result among the methods considered under the matched target construction and metric offline, surpassing both specialized video-grounding systems and strong general-purpose multimodal models.

BaGLM addresses the challenging setting of training-free online step grounding, whereas the proposed method focuses on constructing compact, goal-oriented highlights from long egocentric video. Under the shared Ego4D evaluation protocol, the stronger performance of Visual Self-Alignment likely reflects the combined effect of fine-grained SigLIP 2 frame–step representations, step-wise candidate filtering, and order-aware sequence decoding. These components jointly reduce ambiguity among repeated actions and visually similar steps, producing more accurate step-specific temporal predictions. The results therefore highlight complementary strengths: BaGLM

advances online procedural grounding, while Visual Self-Alignment is particularly effective for extracting compact and temporally organized procedural evidence.

5.3 Component Ablation

To quantify the contribution of the principal design choices, controlled ablations examine the visual–language encoder, sequence-level alignment, and directional transition structure. The two operating modes of the main method, Intention-Primed Alignment and Visual Self-Alignment, additionally reflect the effect of externally supplied temporal guidance. This experiment is conducted on the validation split using identical inputs across all settings.

CLIP Encoder replaces SigLIP 2 with CLIP ViT-L/14 at 336 px while retaining the candidate-selection procedure, temporal prior, prior weight, Gaussian bandwidth, order-aware decoder, and post-decoding coverage rule. This ablation isolates the effect of the visual–language representation.

Independent Visual Ranking removes the temporal alignment and sequence decoding modules. Candidate frames are ranked independently by their maximum frame–step similarity and assigned to the corresponding step. The same primary frame budget and post-selection coverage rule are retained. This ablation quantifies the contribution of structured temporal alignment relative to independent visual–semantic retrieval.

Bidirectional Transitions retains the complete Intention-Primed Alignment pipeline but removes the directional preference among active step states. Forward and backward transitions of the same temporal distance receive identical scores, while the visual emissions, temporal prior, background-state transitions, initialization, and coverage rule remain unchanged. This ablation isolates the contribution of forward-biased procedural ordering while preserving global Viterbi decoding.

Table 4 reports the ablation results at $P = 10\%$. The ablation results reveal a clear hierarchy among the design components. Independent Visual Ranking performs substantially worse than the complete method, showing that semantic similarity provides useful local evidence but is insufficient on its own. Repeated objects, actions, and environments make individual frame–step matches ambiguous in long egocentric video. Sequence-level decoding resolves this ambiguity by evaluating assignments in the context of the surrounding candidates and the full procedural sequence, transforming locally plausible matches into a globally coherent alignment. The encoder ablation further shows that structured decoding depends on the quality of its visual evidence. Replacing SigLIP 2 with CLIP weakens the separation of fine-grained actions, while the much larger degradation from Independent Visual Ranking indicates that stronger representations cannot replace sequence-level reasoning. The smaller effect of Bidirectional Transitions suggests that most of the benefit comes from jointly optimizing assignments across time, with encoder quality and directional structure providing complementary refinement. As a separate diagnostic, Window-Regularized Alignment evaluates the Gaussian temporal prior under oracle parameterization using annotated step intervals, while keeping all other components unchanged. This condition achieves 92.86% Step F1 and 74.74% Frame Precision, indicating that the Gaussian prior can support substantially

Table 4. Component ablation at $P = 10\%$, isolating the contributions of the visual–language encoder, sequence-level alignment, and directional transition structure.

Method	Step Recall (%)	Step F1 (%)	Frame Precision (%)
Intention-Primed Alignment	82.19	82.19	57.22
Bidirectional Transitions	80.71	80.71	55.41
CLIP Encoder	77.59	77.59	50.56
Independent Visual Ranking	50.20	50.20	20.33

Table 5. Stage-wise runtime decomposition of the complete highlight-extraction pipeline. Other pipeline overhead is the difference between the measured pipeline total and the sum of the instrumented stages. The end-to-end total additionally includes per-video model loading (2.77 s) and process setup.

Stage	Mean time (s)	Pipeline share (%)
Full-frame frame-text scoring	71.89	60.10
Candidate selection and I/O	5.24	4.38
Dense candidate rescoring	41.91	35.04
Temporal alignment and sequence decoding	0.29	0.24
Evaluation	0.06	0.05
Other pipeline overhead	0.23	0.19
Pipeline total	119.62	100.00
End-to-end total	123.08	–

stronger alignment when accurately parameterized. These results characterize the potential of the prior rather than a deployable operating point, since the required interval annotations are unavailable at inference time.

5.4 Computational Efficiency

This analysis quantifies the runtime and memory cost of the complete highlight-extraction pipeline, with particular attention to the overhead introduced by temporal alignment and sequence decoding. Intention-Primed Alignment is profiled at $P = 10\%$ on 50 videos. Each video is processed independently on an NVIDIA RTX 6000 Ada GPU using fresh text and image inference, without cached embeddings. The sample has a mean duration of 28.03 minutes and includes videos up to 120.40 minutes long. Runtime is decomposed into vision-language scoring, candidate processing, temporal alignment, and evaluation, while GPU memory is measured for both the complete pipeline and the alignment stage.

The mean pipeline runtime is 119.62 seconds per video. Including per-video model loading, the mean end-to-end runtime is 123.08 seconds per video, equivalent to 263.4 GPU-seconds per hour of source video, $13.67\times$ real-time processing, and a real-time factor of 0.0732. Full-frame scoring and dense candidate rescoring account for 95.14% of the pipeline runtime. Temporal alignment and sequence decoding require only 0.29 seconds per video, corresponding to 0.24% of the total pipeline time. Computation is therefore dominated by pretrained vision-language inference rather than structured sequence optimization.

The memory measurements show the same pattern. The complete pipeline reaches a mean peak of 6.29 GB allocated and 7.89 GB reserved, while temporal alignment adds only 8.16 MB above the loaded-model baseline. For the sampled video closest to the median duration, the Viterbi dynamic-programming table occupies 0.196 MiB. At this operating point, the final highlights retain 1.87% of the corresponding 30 FPS source frames. The pipeline therefore substantially reduces the visual context provided to downstream models, while temporal alignment contributes negligible runtime and memory overhead.

6 DOWNSTREAM APPLICATION: CONTEXTUAL QUESTION ANSWERING

A central challenge for AI-enabled glasses and wearable agents is to transform continuous first-person video into a persistent, queryable memory of everyday activity, enabling questions such as where an object was left or whether a task step was completed. Meeting this challenge requires preserving the visual evidence needed for later retrieval and reasoning while discarding redundant content. Goal-oriented highlight extraction addresses this problem by selecting frames aligned with the user’s intended goal-steps, thereby producing a compact record

of task-relevant evidence. For wearable AI systems, this representation can reduce the storage and transmission burden of continuous video, as well as the visual-token usage, inference latency of downstream multimodal models. The following question-answering evaluation tests whether these efficiency gains can be achieved without sacrificing the information required for episodic recall and long-horizon video understanding.

Although Section 5 evaluates temporal alignment and step coverage, these metrics do not determine whether the selected highlights retain the information required by downstream applications. A temporally accurate highlight sequence may still omit object states, action details, or temporal relations needed to answer a question. We therefore evaluate the extracted highlights through contextual video question answering. A multimodal language model receives either densely sampled frames or the corresponding goal-oriented highlights and answers the same questions under both conditions. The resulting difference in answer quality quantifies how much question-relevant information is lost through highlight extraction.

6.1 Dataset

The downstream evaluation uses MultiHop-EgoQA [8], a long-form egocentric video question-answering benchmark designed for multi-hop reasoning over temporally distributed evidence. Each 180-second video segment is associated with question-answer pairs from six reasoning categories, including repeated activities, event composition, and multiple-object interactions. These questions provide a challenging test of summary quality because answering them often requires evidence from multiple moments across the video. The selected highlights must therefore preserve not only isolated salient events, but also the chain of visual evidence needed to connect them across the activity timeline. MultiHop-EgoQA further challenges goal-oriented extraction because its questions are derived from dense narrations and may target incidental actions or contextual details outside the provided goal-steps. The evaluation thus examines whether goal-oriented highlights retain sufficient temporal coverage and contextual information to support downstream reasoning beyond the content explicitly prioritized during extraction.

6.2 Experiment Setup

The evaluation uses a paired comparison between **Dense Uniform Sampling** and **Goal-Oriented Highlights**. Dense Uniform Sampling extracts one frame per second from each video segment. Goal-Oriented Highlights uses the frames selected at $P = 10\%$. When the same source frame is selected for multiple goal-steps, it is included only once in the chronological input sequence. Different conditions use the same videos, questions, model configuration, image preprocessing, prompting, chunking procedure, and evaluation protocol, and differ only in the frames supplied to the question-answering model. GPT-5.5 serves as the question-answering model. For each video segment, the model receives the timestamped frames from one condition together with all associated questions. Each frame is accompanied only by its timestamp; the model is not given the input condition, goal-step labels, reference answers, or annotated evidence intervals. The model is instructed to answer solely from the supplied visual evidence and to return “Unknown” when the evidence is insufficient. When a segment exceeds the per-request image limit, its frames are divided into consecutive temporal chunks. The complete question set is answered from each chunk, and the chunk-level responses are reconciled in a subsequent text-only call. No frames are discarded during this process.

Answer quality is assessed using a separate GPT-5.5 invocation as an LLM judge. For each question, the judge receives the reference answer and the predictions from different visual-input conditions under anonymous labels. The order of the dense and highlight predictions is balanced across questions, and each prediction is scored independently using the following five-point rubric based on factual correctness and semantic completeness:

- **5 (Excellent):** The prediction is semantically equivalent to the reference answer and accurately captures all essential information.

- **4 (Good):** The prediction is largely correct and captures most essential information, with only a minor omission or discrepancy.
- **3 (Fair):** The prediction is partially correct but contains a notable omission or inaccuracy.
- **2 (Poor):** The prediction contains limited relevant information but includes a major error or omission.
- **1 (Very Poor):** The prediction is incorrect, irrelevant, or fails to provide an answer supported by the reference.

The paired score difference between the two conditions measures the change in question-answering quality when dense uniform sampling is replaced by goal-oriented highlights. A small difference indicates that the highlights retain most of the question-relevant information available under dense sampling.

6.3 Results and Analysis

Table 6 summarizes the trade-off between downstream QA accuracy and the visual context required for multimodal inference. Compared with dense 1 FPS sampling, both highlight-based conditions reduce the number of input frames by approximately 32%, and total API-token usage by approximately 24%. These efficiency gains are accompanied by modest decreases in mean QA score of 0.167 points for Intention-Primed Alignment and 0.193 points for Visual Self-Alignment. Figure 4 provides a qualitative example of how goal-oriented frame selection changes the visual evidence presented to the QA model. For the wooden-spoon question, the highlight condition captures multiple interactions and scores 4, whereas uniform sampling identifies only stirring and scores 2. Both conditions correctly resolve the milk–flour ordering and score 5. These examples show that goal-oriented selection can preserve action and temporal evidence while removing redundant frames, and may occasionally support more complete answers than dense sampling. Taken together, these results indicate that much of the visual evidence needed for downstream QA is concentrated in a relatively small subset of frames, whereas redundant context accounts for a substantial portion of the processing burden.

7 DISCUSSION

7.1 From Video Compression to Goal-Conditioned Visual Memory

This work introduces an intention-driven framework for extracting highlights from long-form egocentric video while preserving the visual evidence required for reasoning and recall. A natural question raised by this design is

Table 6. Comparison among dense 1 FPS sampling, Intention-Primed Alignment (IPA), and Visual Self-Alignment (VSA) for downstream QA at $P = 10\%$. Negative values in the last two columns indicate reductions relative to dense 1 FPS sampling.

Metric	Dense Uniform 1 FPS	Intention-Primed Alignment	Visual Self- Alignment	Δ IPA vs. Dense	Δ VSA vs. Dense
Resource usage					
Total frames sent to QA model	48,240	32,841	32,999	-31.9%	-31.6%
Average frames per video	374.0	254.6	255.8	-31.9%	-31.6%
Average tokens per video	59,265	44,874	45,128	-24.3%	-23.9%
Total tokens	7,645,172	5,788,718	5,821,566	-24.3%	-23.9%
Average latency per video	169.1 s	138.1 s	161.4 s	-18.3%	-4.6%
Average API cost per video	\$0.412	\$0.308	\$0.311	-25.3%	-24.6%
Total API cost	\$53.16	\$39.71	\$40.06	-25.3%	-24.6%
QA quality					
Average question–answer score	2.933	2.766	2.740	-0.167	-0.193

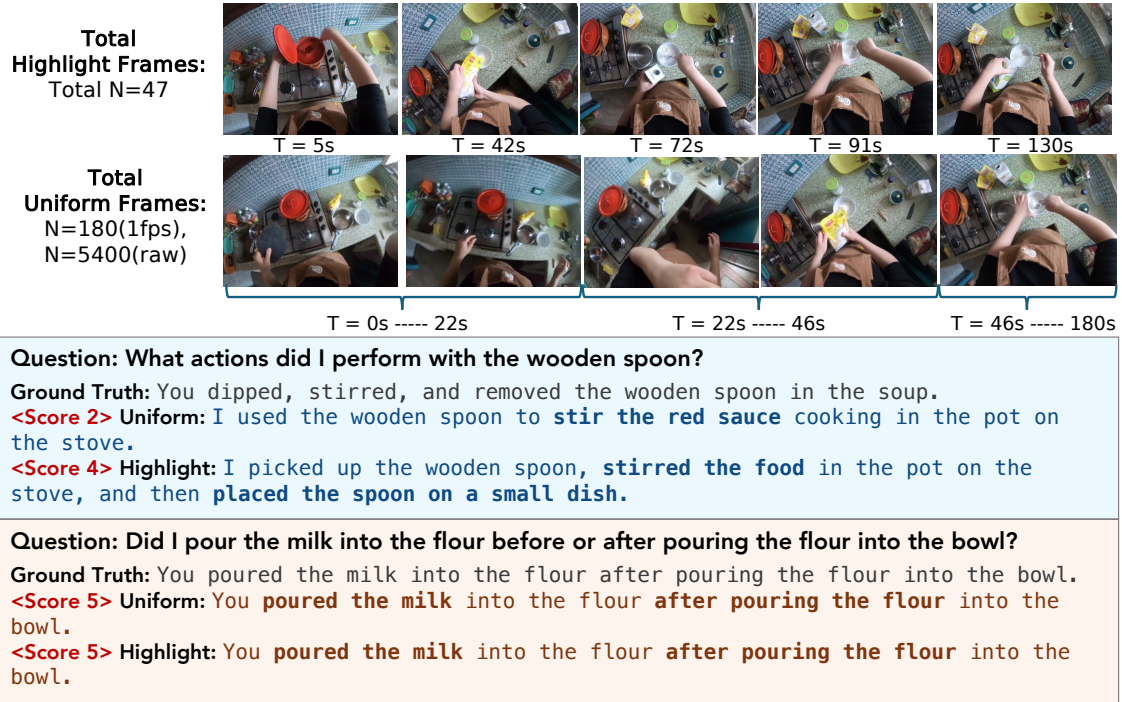


Fig. 4. Qualitative comparison of uniform 1 FPS sampling and goal-oriented highlight extraction for video QA. The highlight method retains 47 frames, compared with 180 frames under uniform sampling. A single frame at $T = 42$ s captures the flour-pouring event, while a sparse set of frames from $T = 46$ – 180 s represents repeated stirring. The highlight input achieves QA scores of 4 and 5, compared with 2 and 5 under uniform sampling, while removing redundant and transitional frames.

whether an explicit goal-step plan is always available in practice. Rather than treating goal-steps as a limitation, this work adopts them as a deliberate abstraction that defines the problem scope and separates highlight extraction from upstream goal and step inference. Given an ordered natural-language description of the intended goal-steps, the proposed method selects a compact, temporally coherent set of visual evidence. Recent work addresses complementary upstream problems, including goal inference, plan recognition, procedural decomposition, and temporal grounding in long-form egocentric video. This paper instead focuses on the subsequent challenge of converting an available procedural description and a noisy, redundant visual stream into a compact representation for downstream retrieval and reasoning. This separation makes the framework modular. Goal-steps may be provided through explicit user input, voice commands, task specifications, known procedures, or automated inference models. Improvements in upstream goal and step inference can be incorporated without modifying the highlight-selection and alignment mechanism. In many practical settings, including instructional activities, planned routines, and agent-assisted tasks, the goal and expected steps are already available. In such cases, the central challenge is not discovering the intention, but managing the scale and redundancy of the recorded visual experience.

More broadly, goal conditioning changes video summarization from producing a single generic synopsis to constructing task-specific visual memories. For wearable devices and agentic AI systems, the relevance of past observations depends on the current question or action. Locating an object, verifying a completed step, and recovering context for a future decision may each require different evidence from the same recording. The

proposed framework supports these uses by mapping a shared egocentric video history to distinct, temporally ordered highlight sets. These compact representations reduce the visual content processed by downstream video-language models while preserving the evidence needed for retrieval and reasoning. Goal-oriented highlight extraction therefore provides a representation layer for transforming continuous egocentric video into compact, query-specific visual evidence that can be stored, retrieved, and analyzed by contextual AI agents.

7.2 Limitations and Future Work

GOHi has several limitations that define directions for future work. First, the framework separates highlight extraction from downstream reasoning. Although this modular design improves efficiency, it creates an irreversible information bottleneck: visual evidence omitted during selection cannot be recovered by the downstream model. This limitation is particularly important when the relevance of a subtle detail becomes apparent only after a query is posed, such as the state of a small component during a repair task. Future work should investigate adaptive selection and reasoning mechanisms that can revisit original videos, retrieve additional evidence, or update retained visual memory when current context is insufficient. Such mechanisms could preserve the efficiency of compact highlight representations while reducing errors caused by premature frame removal.

Second, prioritizing goal-aligned evidence can narrow contextual coverage, and the current decoder does not explicitly reject off-task activity or execution errors. Incidental actions, visually similar distractors, or restarted steps may therefore be assigned to the closest target step, while contextual details that later become relevant may be omitted. Future work should investigate explicit rejection mechanisms, error-aware procedural modeling, and multi-granularity representations that preserve both goal-focused evidence and broader situational context. Future work could also improve robustness to inaccurate recollections through uncertainty-aware temporal priors or explicit temporal-cue correction. Third, the current framework relies on visual input, although audio provides complementary evidence in egocentric recordings. Speech, activity sounds, and ambient events can help infer the wearer’s goal, disambiguate visually similar actions, and signal relevant events that occur outside the camera’s field of view or under occlusion. Future work should integrate synchronized audio. Audio-visual memory representations could therefore improve the coverage of relevant events and support more reliable downstream reasoning, while requiring robust treatment of environmental noise.

Finally, the downstream evaluation measures factual and procedural utility through question answering, but does not capture the experiential or affective qualities of memory. A summary may preserve what occurred while omitting cues related to frustration, satisfaction, or social interaction. These cues may be important in applications centered on reminiscence, reflection, and well-being. Future work should therefore investigate multimodal summarization methods that preserve not only task-relevant information, but also broader affective and social context. Together, these limitations define a broader research agenda that moves beyond efficient video compression toward adaptive, context-sensitive wearable memory systems. By consolidating continuous recordings into compact, retrievable memories, wearable personal AI systems could ground their assistance in the wearer’s lived experience, support everyday recall and task resumption, and help people with memory loss.

8 CONCLUSION

This work presents a training-free framework for goal-oriented highlight extraction from long-form egocentric video. Given an ordered natural-language description of the user’s goal-steps, the method combines open-vocabulary vision–language alignment, adaptive candidate filtering, and global sequence optimization to identify a compact set of frames that preserves procedural evidence. This formulation moves beyond generic saliency and independent frame retrieval by treating the user’s goal as the organizing structure of the visual stream. The evaluation of GOHi on Ego4D Goal-Step shows that structured, goal-conditioned selection provides stronger procedural coverage and frame-level alignment than content-agnostic sampling, generic visual summarization,

independent vision–language retrieval, and multimodal semantic scoring. Under the evaluation protocol used by recent training-free video step grounding, the extracted highlights also achieve the strongest reported result among the compared methods. The component analyses attribute these gains to the combination of strong semantic representations and sequence-level alignment, while computational profiling shows that the structured optimization adds negligible runtime and memory overhead relative to vision–language inference.

The downstream question-answering evaluation further suggests that goal-oriented highlights can reduce the number of frames, token usage, and latency required by a multimodal model while incurring only a modest decrease in answer quality relative to dense sampling. Taken together, these findings indicate that goal-oriented highlight extraction offers a promising mechanism for constructing compact, queryable visual memory from continuous wearable video. More broadly, this work contributes to ongoing efforts to connect long-term perception with contextual and agentic reasoning. A central challenge for such systems is to retain and retrieve relevant prior evidence under limited memory and computational budgets. By organizing extended visual streams around user goals, the proposed framework provides one approach for making past experience more accessible to downstream retrieval, verification, and reasoning. Such representations may support future AI-glasses and wearable agents in maintaining useful context across extended interactions while reducing reliance on exhaustive processing of the original visual stream.

REFERENCES

- [1] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. 2025. Qwen2.5-VL Technical Report. *arXiv preprint arXiv:2502.13923* (2025).
- [2] Mirza Mansoor Baig, Shereen Afifi, Hamid GholamHosseini, and Farhaan Mirza. 2019. A systematic review of wearable sensors and IoT-based monitoring applications for older adults—a focus on ageing population and independent living. *Journal of medical systems* 43, 8 (2019), 233.
- [3] Sven Bambach, Stefan Lee, David J Crandall, and Chen Yu. 2015. Lending a hand: Detecting hands and recognizing activities in complex egocentric interactions. In *Proceedings of the IEEE international conference on computer vision*. 1949–1957.
- [4] Siddhant Bansal, Chetan Arora, and CV Jawahar. 2022. My view is the best view: Procedure learning from egocentric videos. In *European Conference on Computer Vision*. Springer, 657–675.
- [5] Emma Berry, Narinder Kapur, Lyndsay Williams, Steve Hodges, Peter Watson, Gavin Smyth, James Srinivasan, Reg Smith, Barbara Wilson, and Ken Wood. 2007. The use of a wearable camera, SenseCam, as a pictorial diary to improve autobiographical memory in a patient with limbic encephalitis: A preliminary report. *Neuropsychological rehabilitation* 17, 4-5 (2007), 582–601.
- [6] Lucas Beyer, Pavel Izmailov, Alexander Kolesnikov, Mathilde Caron, Simon Kornblith, Xiaohua Zhai, Matthias Minderer, Michael Tschannen, Ibrahim Alabdulmohsin, and Filip Pavetic. 2023. Flexivit: One model for all patch sizes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 14496–14506.
- [7] Chien-Yi Chang, De-An Huang, Yanan Sui, Li Fei-Fei, and Juan Carlos Niebles. 2019. D3tw: Discriminative differentiable dynamic time warping for weakly supervised action alignment and segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 3546–3555.
- [8] Qirui Chen, Shangzhe Di, and Weidi Xie. 2025. Grounded multi-hop videoqa in long-form egocentric videos. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 39. 2159–2167.
- [9] Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. 2024. Expanding Performance Boundaries of Open-Source Multimodal Models with Model, Data, and Test-Time Scaling. *arXiv preprint arXiv:2412.05271* (2024).
- [10] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. 2018. Scaling egocentric vision: The epic-kitchens dataset. In *Proceedings of the European conference on computer vision (ECCV)*. 720–736.
- [11] Sandra Eliza Fontes De Avila, Ana Paula Brandao Lopes, Antonio da Luz Jr, and Arnaldo de Albuquerque Araújo. 2011. VSUMM: A mechanism designed to produce static video summaries and a novel evaluation method. *Pattern recognition letters* 32, 1 (2011), 56–68.
- [12] Mostafa Dehghani, Basil Mustafa, Josip Djolonga, Jonathan Heek, Matthias Minderer, Mathilde Caron, Andreas Steiner, Joan Puigcerver, Robert Geirhos, Ibrahim M Alabdulmohsin, et al. 2023. Patch n’pack: Navit, a vision transformer for any aspect ratio and resolution. *Advances in Neural Information Processing Systems* 36 (2023), 2252–2274.
- [13] Ana Garcia Del Molino, Cheston Tan, Joo-Hwee Lim, and Ah-Hwee Tan. 2016. Summarization of egocentric videos: A comprehensive survey. *IEEE Transactions on Human-Machine Systems* 47, 1 (2016), 65–76.

- [14] Mikita Dvornik, Isma Hadji, Konstantinos G Derpanis, Animesh Garg, and Allan Jepson. 2021. Drop-dtw: Aligning common signal between sequences while dropping outliers. *Advances in Neural Information Processing Systems* 34 (2021), 13782–13793.
- [15] Nikita Dvornik, Isma Hadji, Ran Zhang, Konstantinos G Derpanis, Richard P Wildes, and Allan D Jepson. 2023. Stepformer: Self-supervised step discovery and localization in instructional videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 18952–18961.
- [16] Jiri Fajtl, Hajar Sadeghi Sokeh, Vasileios Argyriou, Dorothy Monekosso, and Paolo Remagnino. 2018. Summarizing videos with attention. In *Asian conference on computer vision*. Springer, 39–54.
- [17] Yue Fan, Xiaojian Ma, Rongpeng Su, Jun Guo, Rujie Wu, Xi Chen, and Qing Li. 2025. Embodied videoagent: Persistent memory from egocentric videos and embodied sensors enables dynamic scene understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 6342–6352.
- [18] Han Fang, Pengfei Xiong, Luhui Xu, and Yu Chen. 2021. Clip2video: Mastering video-text retrieval via image clip. *arXiv preprint arXiv:2106.11097* (2021).
- [19] Alircza Fathi, Jessica K Hodgins, and James M Rehg. 2012. Social interactions: A first-person perspective. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 1226–1233.
- [20] Alireza Fathi, Xiaofeng Ren, and James M Rehg. 2011. Learning to recognize objects in egocentric activities. In *CVPR 2011*. IEEE, 3281–3288.
- [21] Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. 2017. Tall: Temporal activity localization via language query. In *Proceedings of the IEEE international conference on computer vision*. 5267–5275.
- [22] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. 2022. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 18995–19012.
- [23] Kristen Grauman, Andrew Westbury, Lorenzo Torresani, Kris Kitani, Jitendra Malik, Triantafyllos Afouras, Kumar Ashutosh, Vijay Baiyya, Siddhant Bansal, Bikram Boote, et al. 2024. Ego-exo4d: Understanding skilled human activity from first-and third-person perspectives. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 19383–19400.
- [24] Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. 2006. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd international conference on Machine learning*. 369–376.
- [25] Michael Gygli, Helmut Grabner, Hayko Riemenschneider, and Luc Van Gool. 2014. Creating summaries from user videos. In *European conference on computer vision*. Springer, 505–520.
- [26] Michael Gygli, Helmut Grabner, and Luc Van Gool. 2015. Video summarization by learning submodular mixtures of objectives. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 3090–3098.
- [27] Tengda Han, Weidi Xie, and Andrew Zisserman. 2022. Temporal alignment networks for long-term video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2906–2916.
- [28] Ming He, Yong Ge, Enhong Chen, Qi Liu, and Xuesong Wang. 2017. Exploring the emerging type of comment for online videos: Danmu. *ACM Transactions on the Web (TWEB)* 12, 1 (2017), 1–33.
- [29] Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. 2018. Localizing moments in video with temporal language. *arXiv preprint arXiv:1809.01337* (2018).
- [30] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*. PMLR, 4904–4916.
- [31] Ruochen Jiang, Changbo Qu, Jiannan Wang, Chi Wang, and Yudian Zheng. 2020. Towards extracting highlights from recorded live videos: An implicit crowdsourcing approach. In *2020 IEEE 36th International Conference on Data Engineering (ICDE)*. IEEE, 1810–1813.
- [32] Gunhee Kim, Leonid Sigal, and Eric P Xing. 2014. Joint summarization of large-scale collections of web images and videos for storyline reconstruction. In *Proceedings of the IEEE Conference on computer vision and pattern recognition*. 4225–4232.
- [33] Hilde Kuehne, Alexander Richard, and Juergen Gall. 2017. Weakly supervised learning of actions from transcripts. *Computer Vision and Image Understanding* 163 (2017), 78–89.
- [34] Yong Jae Lee, Joydeep Ghosh, and Kristen Grauman. 2012. Discovering important people and objects for egocentric video summarization. In *2012 IEEE conference on computer vision and pattern recognition*. IEEE, 1346–1353.
- [35] Jie Lei, Tamara L Berg, and Mohit Bansal. 2021. Detecting moments and highlights in videos via natural language queries. *Advances in Neural Information Processing Systems* 34 (2021), 11846–11858.
- [36] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems* 33 (2020), 9459–9474.
- [37] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. 2025. LLaVA-OneVision: Easy Visual Task Transfer. *Transactions on Machine Learning Research* (2025).

- [38] Jun Li, Peng Lei, and Sinisa Todorovic. 2019. Weakly supervised energy-based learning for action segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*. 6243–6251.
- [39] Xiang Li, Heqian Qiu, Lanxiao Wang, Hanwen Zhang, Chenghao Qi, Linfeng Han, Huiyu Xiong, and Hongliang Li. 2025. Challenges and Trends in Egocentric Vision: A Survey. *arXiv preprint arXiv:2503.15275* (2025).
- [40] Zeqian Li, Qirui Chen, Tengda Han, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2024. Multi-sentence grounding for long-term instructional video. In *European Conference on Computer Vision*. Springer, 200–216.
- [41] Kevin Qinghong Lin, Jinpeng Wang, Mattia Soldan, Michael Wray, Rui Yan, Eric Z Xu, Difei Gao, Rong-Cheng Tu, Wenzhe Zhao, Weijie Kong, et al. 2022. Egocentric video-language pretraining. *Advances in Neural Information Processing Systems* 35 (2022), 7575–7586.
- [42] Zheng Lu and Kristen Grauman. 2013. Story-driven summarization for egocentric video. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2714–2721.
- [43] Huaishao Luo, Lei Ji, Ming Zhong, Yang Chen, Wen Lei, Nan Duan, and Tianrui Li. 2021. Clip4clip: An empirical study of clip for end to end video clip retrieval. *arXiv preprint arXiv:2104.08860* (2021).
- [44] Behrooz Mahasseni, Michael Lam, and Sinisa Todorovic. 2017. Unsupervised Video Summarization With Adversarial LSTM Networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- [45] Arjun Majumdar, Anurag Ajay, Xiaohan Zhang, Pranav Putta, Sriram Yenamandra, Mikael Henaff, Sneha Silwal, Paul Mcvay, Oleksandr Maksymets, Sergio Arnaud, et al. 2024. Openqa: Embodied question answering in the era of foundation models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 16488–16498.
- [46] Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. 2023. Egoschema: A diagnostic benchmark for very long-form video language understanding. *Advances in Neural Information Processing Systems* 36 (2023), 46212–46244.
- [47] Kevis-Kokitsi Maninis, Kaifeng Chen, Soham Ghosh, Arjun Karpur, Koert Chen, Ye Xia, Bingyi Cao, Daniel Salz, Guangxing Han, Jan Dlabal, et al. 2024. TIPS: Text-image pretraining with spatial awareness. *arXiv preprint arXiv:2410.16512* (2024).
- [48] Effrosyni Mavroudi, Triantafyllos Afouras, and Lorenzo Torresani. 2023. Learning to ground instructional articles in videos through narrations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 15201–15213.
- [49] Muhammad Ferjad Naeem, Yongqin Xian, Xiaohua Zhai, Lukas Hoyer, Luc Van Gool, and Federico Tombari. 2024. Silc: Improving vision language pretraining with self-distillation. In *European Conference on Computer Vision*. Springer, 38–55.
- [50] Medhini Narasimhan, Anna Rohrbach, and Trevor Darrell. 2021. Clip-it! language-guided video summarization. *Advances in neural information processing systems* 34 (2021), 13988–14000.
- [51] Adrián Núñez-Marcos, Gorka Azkune, and Ignacio Arganda-Carreras. 2022. Egocentric vision-based action recognition: A survey. *Neurocomputing* 472 (2022), 175–197.
- [52] Wei-Ting Peng, Wei-Ta Chu, Chia-Han Chang, Chien-Nan Chou, Wei-Jia Huang, Wen-Yan Chang, and Yi-Ping Hung. 2011. Editing by viewing: automatic home video summarization by viewing behavior analysis. *IEEE Transactions on Multimedia* 13, 3 (2011), 539–550.
- [53] Hamed Pirsiavash and Deva Ramanan. 2012. Detecting activities of daily living in first-person camera views. In *2012 IEEE conference on computer vision and pattern recognition*. IEEE, 2847–2854.
- [54] Danila Potapov, Matthijs Douze, Zaid Harchaoui, and Cordelia Schmid. 2014. Category-specific video summarization. In *European conference on computer vision*. Springer, 540–555.
- [55] Shraman Pramanick, Yale Song, Sayan Nag, Kevin Qinghong Lin, Hardik Shah, Mike Zheng Shou, Rama Chellappa, and Pengchuan Zhang. 2023. Egovlpv2: Egocentric video-language pre-training with fusion in the backbone. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, 5262–5274.
- [56] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*. PmLR, 8748–8763.
- [57] Francesco Ragusa, Antonino Furnari, Salvatore Livatino, and Giovanni Maria Farinella. 2021. The meccano dataset: Understanding human-object interactions from egocentric videos in an industrial-like domain. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 1569–1578.
- [58] Shuhuai Ren, Linli Yao, Shicheng Li, Xu Sun, and Lu Hou. 2024. Timechat: A time-sensitive multimodal large language model for long video understanding. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 14313–14323.
- [59] Xubin Ren, Lingrui Xu, Long Xia, Shuaiqiang Wang, Dawei Yin, and Chao Huang. 2026. Videorag: Retrieval-augmented generation with extreme long-context videos. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 1*. 2390–2401.
- [60] Alexander Richard, Hilde Kuehne, Ahsan Iqbal, and Juergen Gall. 2018. Neuralnetwork-viterbi: A framework for weakly supervised video learning. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*. 7386–7395.
- [61] Ivan Rodin, Antonino Furnari, Dimitrios Mavroeidis, and Giovanni Maria Farinella. 2021. Predicting the future from first person (egocentric) vision: A survey. *Computer Vision and Image Understanding* 211 (2021), 103252.
- [62] Hiroaki Sakoe and Seibi Chiba. 1978. Dynamic programming algorithm optimization for spoken word recognition. *IEEE transactions on acoustics, speech, and signal processing* 26, 1 (1978), 43–49.

- [63] Fadime Sener, Dibiyadip Chatterjee, Daniel Shelepov, Kun He, Dipika Singhania, Robert Wang, and Angela Yao. 2022. Assembly101: A large-scale multi-view video dataset for understanding procedural activities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 21096–21106.
- [64] Aidean Sharghi, Boqing Gong, and Mubarak Shah. 2016. Query-focused extractive video summarization. In *European conference on computer vision*. Springer, 3–19.
- [65] Aidean Sharghi, Jacob S Laurel, and Boqing Gong. 2017. Query-focused video summarization: Dataset, evaluation, and a memory network based approach. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 4788–4797.
- [66] Xiaoqian Shen, Yunyang Xiong, Changsheng Zhao, Lemeng Wu, Jun Chen, Chenchen Zhu, Zechun Liu, Fanyi Xiao, Balakrishnan Varadarajan, Florian Bordes, et al. 2024. Longvu: Spatiotemporal adaptive compression for long video-language understanding. *arXiv preprint arXiv:2410.17434* (2024).
- [67] Yale Song, Eugene Byrne, Tushar Nagarajan, Huiyu Wang, Miguel Martin, and Lorenzo Torresani. 2023. Ego4d goal-step: Toward hierarchical understanding of procedural activities. *Advances in Neural Information Processing Systems* 36 (2023), 38863–38886.
- [68] Jinhwan Sul, Jihoon Han, and Joonseok Lee. 2023. Mr. hisum: A large-scale dataset for video highlight detection and summarization. *Advances in Neural Information Processing Systems* 36 (2023), 40542–40555.
- [69] Min Sun, Ali Farhadi, and Steve Seitz. 2014. Ranking domain-specific highlights by analyzing edited videos. In *European conference on computer vision*. Springer, 787–802.
- [70] Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. 2024. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295* (2024).
- [71] Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, et al. 2025. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786* (2025).
- [72] Bo Wan, Michael Tschannen, Yongqin Xian, Filip Pavetic, Ibrahim M Alabdulmohsin, Xiao Wang, André Susano Pinto, Andreas Steiner, Lucas Beyer, and Xiaohua Zhai. 2024. Locca: Visual pretraining with location-aware captioners. *Advances in Neural Information Processing Systems* 37 (2024), 116355–116387.
- [73] Mengmeng Wang, Jiazheng Xing, and Yong Liu. 2021. Actionclip: A new paradigm for video action recognition. *arXiv preprint arXiv:2109.08472* (2021).
- [74] Weihang Wang, Zehai He, Wenyi Hong, Yean Cheng, Xiaohan Zhang, Ji Qi, Ming Ding, Xiaotao Gu, Shiyu Huang, Bin Xu, et al. 2025. Lvbench: An extreme long video understanding benchmark. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 22958–22967.
- [75] Guande Wu, Jianzhe Lin, and Claudio T Silva. 2022. Intentvizer: Towards generic query guided interactive video summarization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 10503–10512.
- [76] Shuwen Xiao, Zhou Zhao, Zijian Zhang, Xiaohui Yan, and Min Yang. 2020. Convolutional hierarchical attention network for query-focused video summarization. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 34. 12426–12433.
- [77] Jia Xu, Lopamudra Mukherjee, Yin Li, Jamieson Warner, James M Rehg, and Vikas Singh. 2015. Gaze-enabled egocentric video summarization via constrained submodular maximization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2235–2244.
- [78] Karmesh Yadav, Yusuf Ali, Gunshi Gupta, Yarin Gal, and Zsolt Kira. 2025. FindingDory: A Benchmark to Evaluate Memory in Embodied Agents. *arXiv preprint arXiv:2506.15635* (2025).
- [79] Jingkan Yang, Shuai Liu, Hongming Guo, Yuhao Dong, Xiamengwei Zhang, Sicheng Zhang, Pengyun Wang, Zitang Zhou, Binzhu Xie, Ziyue Wang, et al. 2025. Egolife: Towards egocentric life assistant. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 28885–28900.
- [80] Wenmian Yang, Na Ruan, Wenyuan Gao, Kun Wang, Wensheng Ran, and Weijia Jia. 2017. Crowdsourced time-sync video tagging using semantic association graph. In *2017 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 547–552.
- [81] Hanrong Ye, Haotian Zhang, Erik Daxberger, Lin Chen, Zongyu Lin, Yanghao Li, Bowen Zhang, Haoxuan You, Dan Xu, Zhe Gan, et al. 2025. MMEgo: Towards building egocentric multimodal LLMs for video QA. In *International Conference on Learning Representations*, Vol. 2025. 71705–71723.
- [82] Luca Zanella, Massimiliano Mancini, Yiming Wang, Alessio Tonioni, and Elisa Ricci. 2025. Training-free Online Video Step Grounding. *Advances in Neural Information Processing Systems* 38 (2025), 99372–99398.
- [83] Xiaohua Zhai, Alexander Kolesnikov, Neil Houlsby, and Lucas Beyer. 2022. Scaling vision transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 12104–12113.
- [84] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*. 11975–11986.
- [85] Hao Zhang, Aixin Sun, Wei Jing, Liangli Zhen, Joey Tianyi Zhou, and Rick Siow Mong Goh. 2021. Natural language video localization: A revisit in span-based question answering framework. *IEEE transactions on pattern analysis and machine intelligence* 44, 8 (2021),

4252–4266.

- [86] Ke Zhang, Wei-Lun Chao, Fei Sha, and Kristen Grauman. 2016. Video summarization with long short-term memory. In *European conference on computer vision*. Springer, 766–782.
- [87] Peiyuan Zhang, Kaichen Zhang, Bo Li, Guangtao Zeng, Jingkan Yang, Yuanhan Zhang, Ziyue Wang, Haoran Tan, Chunyuan Li, and Ziwei Liu. 2024. Long context transfer from language to vision. *arXiv preprint arXiv:2406.16852* (2024).
- [88] Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Yuchen Duan, Hao Tian, Weijie Su, Jie Shao, et al. 2025. InternVL3: Exploring Advanced Training and Test-Time Recipes for Open-Source Multimodal Models. *arXiv preprint arXiv:2504.10479* (2025).
- [89] Dimitri Zhukov, Jean-Baptiste Alayrac, Ramazan Gokberk Cinbis, David Fouhey, Ivan Laptev, and Josef Sivic. 2019. Cross-task weakly supervised learning from instructional videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 3537–3545.
- [90] Wazeer Zulfikar, Samantha Chan, and Pattie Maes. 2024. Memoro: Using large language models to realize a concise interface for real-time memory augmentation. *arXiv preprint arXiv:2403.02135* (2024).