EatVid-Bench: A Multimodal Fine-Grained Eating Behavior Video Dataset

Xiangpeng Zheng[1], Zhenbo Xu[3,4], Gong Huang[2], Zhu Li[5], and Qinghong Yang[1,2⋆]

[1] Hangzhou International Innovation Institute of Beihang University, Beihang University, Hangzhou, China
[2] Beihang University, Beijing, China
[3] School of Artificial Intelligence, Beijing University of Posts and Telecommunications, Beijing, China
[4] ShiFang Technology Inc., Hangzhou, China
[5] Department of Mathematics, Imperial College London, London, United Kingdom

Abstract. Understanding fine-grained eating behaviors in unconstrained daily-life videos is essential for dietary monitoring and behavioral health assessment, yet existing video-language benchmarks lack the temporal granularity and domain-specific grounding required to rigorously evaluate models in this health-critical domain. We introduce EatVid-Bench, a large-scale, multi-dimensional benchmark designed to advance fine-grained eating behavior understanding. Our dataset comprises nearly 700 real-world eating sessions (around 3,000 minutes) annotated through a novel three-tier automated pipeline that integrates 12 complementary signals, including body pose, food detection, bite events, and facial expressions, with each annotation traceable to its source signal for full interpretability and verifiability. Building on this foundation, we construct a question-answering benchmark spanning seven capability dimensions, three difficulty levels, and five question types, with rigorous provenance tracking throughout. Comprehensive evaluation of state-of-the-art Video-LLMs exposes a substantial perception–reasoning gap, particularly in sub-second temporal grounding and cross-frame counting. To further validate the effectiveness of EatVid-Bench, we propose a domain-adapted fine-tuning strategy to provide a strong open-source baseline by leveraging structured training annotations as explicit reasoning supervision. Code and benchmark are publicly available at GitHub and HuggingFace.

Keywords: Eating Behavior Understanding · Temporal Video Understanding · Video-Language Benchmark · Multimodal Large Language Models

1 Introduction

Understanding eating behavior is pivotal to human health: it is closely associated with obesity, eating disorders, and metabolic disease [11, 33], and encodes a rich

⋆ Corresponding author.

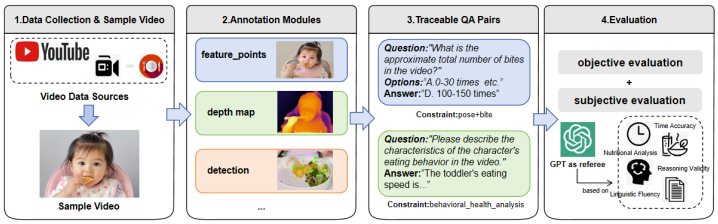

Fig. 1: The EatVid-Bench Dataset. Our benchmark comprises around 3,000 minutes of densely annotated eating videos and nearly 8,000 questions.

behavioral signal manifested through action dynamics, facial expressions, and fine-grained temporal sequences that reflects affective states, dietary habits, and psychological conditions [11, 21, 64]. In clinical and nutritional research, monitoring eating sessions in unconstrained, daily-life settings is essential for delivering personalized dietary interventions and behavioral health assessment [36]. Yet despite the emergence of powerful Video Large Language Models (Video-LLMs) [31, 50] that have elevated video understanding well beyond simple action recognition [57, 59], the eating-behavior domain remains a largely unsolved frontier for these models.

The core difficulty is that meal analysis is not reducible to detecting that someone is eating. It requires integrating fine-grained temporal perception, such as micro-motions in food-to-mouth transfer, with domain-specific discrimination, such as mapping food categories to nutritional groups, and higher-level behavioral reasoning, such as relating eating pace to satiety cues or stress-related signals [39, 48, 55]. Applying general-purpose Video-LLMs to this domain reveals a pronounced capability gap [6, 61]: existing video-language benchmarks mainly target broad daily activities or cinematic scenes, and therefore lack the temporal granularity, health-oriented grounding, and diagnostic rigor needed for eating-behavior assessment [20, 26]. Consequently, coarse action-recognition or video-QA scores may conceal whether a model reasons from physical evidence or merely relies on linguistic priors.

Developing a robust benchmark for eating-behavior understanding therefore poses several interconnected challenges.

Subtle and repetitive temporal grounding. Eating consists of dense micro-actions, including food-to-mouth transfers, chewing, and intermittent distractions such as checking a phone, that are visually subtle and highly repetitive [4, 44, 55]. Models require fine temporal precision to localize events correctly and avoid missing short actions or hallucinating nonexistent ones. No existing video-language benchmark probes this sub-second granularity in a health-oriented context.

Partial domain-knowledge integration. Practical dietary and behavioral assessment depends on more than generic object recognition. Models should incorporate partial, domain-relevant signals, such as mapping fine-grained food categories to coarse nutritional groups [19, 32] and using behavioral proxies (e.g.,

bite frequency, hand-to-mouth trajectories, or depth-derived volume cues) to estimate whether intake patterns fall within expected ranges. This targeted domain knowledge suffices for many behavior- and mental-health inferences while remaining realistic given current annotation granularity [27]. Grounding such knowledge in verifiable, traceable annotations is a prerequisite for trustworthy evaluation.

Complex multi-step reasoning chains. Behavioral health assessment requires synthesizing multiple low-level signals, including bite cadence, affective facial expressions, pauses, and total meal duration, into higher-level conclusions about satiety, stress, or other psychological states [35]. This demands multi-hop, temporally grounded reasoning rather than single-shot classification [4, 44]. Benchmarks that evaluate only final answers cannot distinguish genuine comprehension from plausible-sounding guesses.

Extended temporal contexts. Real-world eating sessions frequently span several minutes to tens of minutes, exceeding the effective context window of many contemporary Video-LLMs [22]. Benchmarks must support long-form video processing and global aggregation of evidence across extended temporal windows [44], a requirement that existing short-clip benchmarks cannot satisfy.

To systematically diagnose the performance boundaries and failure mechanisms of Video-LLMs in eating-behavior understanding, we propose **EatVid-Bench**, as illustrated in Fig. 1. Grounded in the need for provenance-linked, hierarchically structured evaluation, we organize our benchmark around a three-tier automated annotation pipeline that transforms raw video signals into traceable behavioral representations, directly addressing the opacity and shallow granularity of existing approaches. Built on this foundation, EatVid-Bench spans seven capability dimensions and three difficulty levels, progressing from atomic-perception queries to compositional behavioral-health synthesis and enabling stratified diagnosis of where and why model capabilities break down. To support principled comparison across question types and difficulty levels, we introduce a difficulty-weighted composite metric that jointly accounts for temporal grounding accuracy, reasoning validity, and nutritional completeness. We further conduct systematic evaluation of representative Video-LLMs and propose **Annotation-Guided Chain-of-Thought LoRA (AGCoT-LoRA)**, a domain-adapted fine-tuning strategy that leverages structured annotations as explicit reasoning supervision.

Our experimental findings reveal consistent and telling patterns. Models generally handle coarse food-presence detection and single-frame recognition, but performance degrades sharply when tasks require sub-second bite localization, cross-frame counting, or integration of behavioral signals into health-oriented conclusions. A particularly revealing failure mode emerges at Level-3 compositional tasks: several models achieve superficially plausible short-answer scores by drawing on linguistic priors rather than frame-level evidence, a hallucination pattern that coarser benchmarks would entirely miss. These findings expose a fundamental gap between perceptual fluency and genuine behavioral reasoning

in current Video-LLMs, and chart a clear path toward evidence-grounded video understanding for health applications.

Our main contributions are summarized as follows:

- **A large-scale, fine-grained dataset with traceable annotations.** We introduce EatVid-Bench, comprising nearly 700 real-world eating sessions (around 3,000 minutes) annotated across 12 complementary signal dimensions. We design a scalable, fully automated annotation pipeline that ensures each label is traceable to its originating signal, improving interpretability and verifiability.
- **A multi-dimensional QA benchmark.** We present a principled evaluation protocol spanning 7 capability categories, 3 difficulty levels, and 5 question types; each ground-truth answer is accompanied by provenance metadata to serve as a rigorous diagnostic for Video-LLMs.
- **Comprehensive analysis and a domain-adapted baseline.** We perform extensive experiments on representative Video-LLMs and identify recurrent failure modes in temporal localization, long-range context aggregation, and health-related reasoning. To further validate the effectiveness of EatVid-Bench, we propose AGCoT-LoRA, which uses annotation-derived reasoning chains as supervision combined with parameter-efficient LoRA fine-tuning to enhance domain-specific reasoning, establishing a strong open-source baseline.

2 Related Work

2.1 Video-Language Benchmarks

The landscape of video-language understanding has been revolutionized by comprehensive benchmarks [5,24,38,60]. General-purpose benchmarks such as Video-MME, MLVU, and MVBench evaluate Video-LLMs across broad categories like perception and reasoning [15, 24, 68]. However, these datasets primarily focus on diverse, domain-agnostic scenes and lack the fine-grained temporal granularity required for health-centric behavior analysis. Egocentric datasets like Ego4D and Epic-Kitchens capture "in-the-wild" kitchen activities, yet their annotations are centered on object manipulation and action verbs rather than behavioral health metrics [12, 13, 16]. Recently, temporal-centric benchmarks such as E.T. Bench and EgoTempo have begun to emphasize precise event localization [29, 40]. In contrast to these works, EatVid-Bench is the first to bridge the gap between long-form video understanding and domain-specific eating behavior analysis, offering unparalleled annotation density in a health-critical context.

Tab. 1 further positions EatVid-Bench against prior eating-behavior datasets and general video-language benchmarks. Prior eating datasets provide valuable sensor- or action-level supervision, but they do not jointly support provenance-linked video-QA and behavioral-health reasoning; general video-QA datasets provide broad semantic evaluation, but lack eating-specific temporal signals and dietary grounding.

Table 1: Comparison with prior datasets and video-language benchmarks.

Dataset	Hours	QA pairs	Eating	Fine-temp.	Food/health	Video-QA	Provenance
OREBA	∼100h	–	✓	✓	intake	–	–
EatSense	∼50h	–	✓	✓	action/QoM	–	partial
Ego4D / EPIC	∼3000h	partial	partial	✓	–	–	partial
Video-MME / MLVU	∼250h	✓	–	partial	–	✓	–
EatVid-Bench	∼50h	8.2K	✓	✓	✓	✓	✓

2.2 Eating Behavior Analysis

Traditionally, eating behavior research has relied on sensor-based approaches (e.g., OREBA), which utilize IMUs or wrist-worn accelerometers to detect bite counts [23, 45]. While effective for simple event counting, these methods lack visual context and cannot support complex semantic reasoning. Vision-based methods have evolved from specialized gesture recognition (e.g., EatSense to broader intake analysis [52]. However, existing datasets are often limited in scale, restricted to laboratory settings, or focused solely on single-task classification (e.g., "eating" vs. "not eating"). EatVid-Bench moves beyond these limitations by providing (a) large-scale, unconstrained real-world videos, (b) multi-dimensional annotations spanning 12 signals, and (c) a comprehensive Video-QA framework that enables high-level behavioral health evaluation.

2.3 Video-Language Large Models

The advent of Multimodal Large Language Models (MLLMs) has significantly advanced video reasoning [67, 69]. Models like InternVL2 [51], Qwen2-VL [56], and LLaVA-Video [66] have demonstrated strong zero-shot capabilities by integrating powerful vision encoders with large-scale LLMs. Specialized architectures like HumanOmni further incorporate human-centric priors. Despite their prowess, these "generalist" models often struggle with domain-specific grounding—such as precisely estimating eating pace or assessing dietary composition from subtle visual cues. Our work evaluates these state-of-the-art models on a specialized domain and proposes AGCoT-LoRA to address their reasoning deficiencies through structured, annotation-guided fine-tuning.

2.4 Automated Video Annotation Pipelines

High-quality video annotation is notoriously labor-intensive; for instance, Ego4D required tens of thousands of human-hours [16]. To achieve scalability, researchers have increasingly turned to automated modular pipelines [54]. Previous tools have leveraged specialized models like SAM for segmentation [43], and MediaPipe for pose estimation [30]. Recent efforts also utilize Whisper for audio transcription to assist semantic labeling [41]. Our pipeline extends this paradigm by integrating 12 heterogeneous modules into a unified, traceable framework. Unlike prior "black-box" annotation efforts, we prioritize provenance tracking, ensuring that every ground-truth answer in our benchmark can be traced back to its underlying physical signal (e.g., a specific bite detection or pose sequence), thereby enhancing the interpretability of the evaluation process.

Table 2: EatVid-Bench statistics and difficulty anchors.

Aspect	Statistics
Data scale	∼2,000 raw sessions → 856 valid sessions → 682 videos / 8,184 QA pairs. Average length: 4.3 minutes/video.
QA coverage	**Types:** SC 43.8%, FB 23.8%, SA 16.2%, MC 8.1%, TF 8.1%. **Categories:** Composition 24.3%, Bite 22.7%, Timeline 16.2%, Health 16.2%, Pace 12.4%, Facial 8.1%.
Diversity	**Utensils:** chopsticks 42%, spoons 34%, forks 19%, hands 5%. **Scenes:** home 60%, cafeteria/restaurant 30%, others 10%. **Foods:** 9 coarse classes with subcategories.
Anchors	**Chance:** 22.0 Non-SA-WAcc. **Text-only:** 58.2 SA-WAcc. **Human (n=30, k=3):** 72.0 WAcc / 88.0 Non-SA-WAcc.

3 The EatVid-Bench Dataset & Pipeline

3.1 Data Collection and Statistics

Data acquisition. We curated EatVid-Bench to capture the complexity of eating behaviors in naturalistic settings. The dataset contains nearly 700 high-resolution video sessions totaling close to 3,000 minutes, gathered from two primary sources: (1) curated web-based repositories (e.g., YouTube vlogs and educational nutrition content) and (2) voluntarily recorded real-world eating sessions in everyday environments. The average session length is approximately 4.3 minutes; individual sessions range from around 2 minutes to over 20 minutes, providing both short-form and long-form examples for temporal analysis.

Selection criteria. We selected videos to maximize the clarity and completeness of eating episodes. Inclusion criteria required that the footage predominantly frame a main participant, contain clearly visible facial regions and food/hand interactions, and, where possible, capture an entire eating episode from beginning to end (e.g., preparation or plating → consumption → brief post-meal period). These criteria prioritize ecological validity while ensuring the visual evidence needed for fine-grained temporal annotation.

Environments, utensils, and dietary variety. The dataset is drawn mainly from home and cafeteria scenes. Utensils observed include chopsticks, forks, and spoons. We use a two-level food taxonomy: 9 coarse categories plus many fine-grained subcategories to maximize coverage. If an item cannot be identified at fine granularity, it is assigned to its high-level category (e.g., unrecognizable water → "beverages").

3.2 The Automated Annotation Pipeline

Annotating around 3,000 minutes of video across 12 behavioral dimensions through manual labeling would be prohibitively expensive and difficult to standardize. We

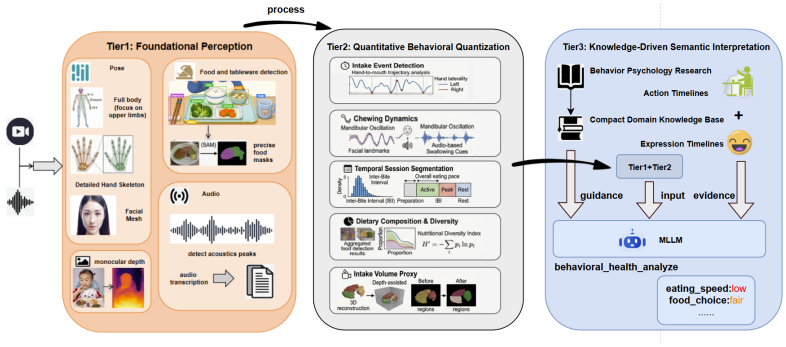

Fig. 2: The EatVid-Bench Automated Annotation Pipeline. The framework employs a hierarchical three-tier architecture to translate raw multi-modal signals into structured behavioral intelligence: Tier 1 extracts low-level primitives including 3D pose, food segmentation, and depth; Tier 2 quantifies domain-specific metrics such as bite events, chewing frequency, and eating pace; Tier 3 leverages MLLMs to generate high-level affective sequences and comprehensive behavioral health reports.

therefore develop a scalable, three-tier automated pipeline (as shown in Fig. 2) that transforms raw video signals into structured, traceable behavioral representations.

Tier 1: Foundational Perception The first tier extracts low-level geometric, semantic, and multimodal primitives from raw pixels.

We employ MediaPipe [9, 30] to extract full-body 3D pose (33 keypoints), detailed hand skeletons (21 keypoints per hand), and dense facial landmarks (468 points). These signals form the basis for trajectory-based motion modeling and fine-grained facial dynamics analysis [9]. We then employ Grounding DINO for food instance detection and category grounding [28], refined with the Segment Anything Model (SAM) for precise segmentation [43]. This produces temporally consistent food masks and category labels. Monocular depth is estimated using MiDaS to provide coarse 3D spatial cues for intake estimation [42]. Ambient audio is transcribed with Whisper-v3 and synchronized with visual streams, enabling detection of swallowing-related acoustic peaks and contextual verbal cues [41].

Tier 2: Quantitative Behavioral Quantization The second tier transforms raw perception signals into domain-specific behavioral metrics.

We detect intake events through hand-to-mouth trajectory analysis [53]. Let d(t) denote the Euclidean distance between the wrist keypoint and the nose-tip landmark at time t. A bite event is anchored at timestamps corresponding to local minima of d(t) below an adaptive threshold. Each detected event is time-stamped and annotated with hand laterality (left/right), enabling subsequent phase analysis.

Chewing dynamics are estimated via mandibular oscillation derived from facial landmarks, fused with audio-based swallowing cues [52]. Temporal synchronization between jaw-motion frequency and acoustic peaks improves robustness under noisy conditions [37].

We compute the Inter-Bite Interval (IBI) distribution and derive the overall eating pace [1,47]. Sessions are further segmented into preparation, active, peak, and rest phases based on bite density statistics.

Food detection results are aggregated over time to estimate dietary composition. We compute a Nutritional Diversity Index using Shannon entropy:

$$H' = -\sum_i p_i \ln p_i \tag{1}$$

where p_i represents the temporal proportion of the i-th food category.

We approximate intake volume using depth-assisted 3D reconstruction and before/after comparisons of segmented food regions, providing a coarse but informative proxy for consumption magnitude.

Tier 3: Knowledge-Driven Semantic Interpretation The third tier bridges quantitative metrics and higher-level reasoning via knowledge-constrained multimodal language models.

Rather than performing unrestricted generative inference, we construct a compact domain knowledge base derived from behavioral psychology research [2, 14, 17, 18, 34, 46, 49]. This knowledge base encodes structured relationships between measurable signals and higher-level interpretations. To capture nuanced behavioral cues, we employ high-capacity MLLMs to generate initial temporal segments for specific actions and facial expressions [62]. During annotation, MLLMs act as "semantic integrators," grounding all conclusions in Tier 1–2 signals, the verified behavioral timelines, and the domain knowledge base. This reduces speculative reasoning and enforces consistency with established principles. Each assertion is automatically tagged with Provenance Metadata (e.g., "Elevated eating pace inferred from average IBI $<$ 15s"), ensuring that all high-level annotations remain auditable and verifiable.

3.3 Annotation Quality and Validation

We enforce a rigorous validation protocol to ensure annotation accuracy and auditability. Human review is prioritized by risk: all Tier-2 and Tier-3 annotations used in evaluation splits were manually examined and corrected. Moreover, Tier-3 semantic outputs are structurally parsed and re-evaluated by a secondary LLM (GPT-4o) to catch language-level hallucinations.

We further validate annotation reliability through provenance-level audits and inter-annotator agreement studies. Tier-1 predictions are not blindly treated as ground truth: we audit source-level signals when they affect QA construction and correct inconsistent provenance traces. For Tier-2 and Tier-3 annotations,

three trained annotators perform double-blind review with third-annotator arbitration, achieving $\kappa = 0.85$ for Tier-2 and $\kappa = 0.82$ for Tier-3. For Short-Answer evaluation, we re-score responses using Gemini, Claude, and human raters, obtaining strong agreement with GPT-4o (Spearman $\rho = 0.89/0.87/0.82$), while Non-SA questions are evaluated with exact matching, regular expressions, or F1 and are independent of LLM judges.

4 The EatVid-QA Benchmark

The core objective of EatVid-QA is to transform raw, multi-modal annotations into a rigorous, verifiable testbed for Video-LLMs. Unlike existing benchmarks that rely on subjective human-written questions, our benchmark is built on the principle of algorithmic provenance, where every question is anchored to a specific physical signal.

4.1 Taxonomy of Eating Behavior Understanding

We categorize the requirements for comprehensive eating behavior understanding into seven functional dimensions, spanning from low-level perception to high-level clinical reasoning.

Fine-Grained Perception: Focuses on Bite Detection (temporal grounding of intake events) and Chewing Analysis (fine-grained motion cycle recognition).

Contextual Semantic Understanding: Includes Temporal Action (narrative timelines), Facial Expression (affective response to food), and Dietary Composition (multi-label food recognition and diversity).

High-Level Behavioral Reasoning: Addresses Eating Pace (statistical aggregation of intervals) and Behavioral Health (comprehensive assessment of eating habits and nutritional balance).

To ensure a robust evaluation, these dimensions are manifested through five distinct question formats: Single Choice (SC), True/False (TF), and Multiple Choice (MC) for objective verification; Fill-in-the-Blank (FB) for precise numerical/categorical extraction; and Short Answer (SA) for open-ended synthesis.

4.2 Traceable QA Generation and Difficulty Scaling

A distinguishing feature of EatVid-QA is its programmatic provenance: question–answer pairs are generated by a template engine that queries Tier-1–3 annotation modules and maps numerical outputs (e.g., pace_stats → "12 bites/min") to objective categorical labels (e.g., "Tachyphagia / Fast"), ensuring that ground truth is reproducible and verifiable.

To probe the limits of Video-LLMs we organize QA pairs by perceptual granularity into a three-tier hierarchy:

Level 1 (Atomic Perception): Identify a single attribute or short-window action (e.g., "What utensil is the subject using at 00:45?").

Level 2 (Temporal Aggregation): Integrate evidence across multiple timestamps (e.g., "How many bite events occurred between the 1st and 3rd minute?").

Level 3 (Compositional Synthesis): Compose signals across modalities and longer windows to produce a higher-level judgment (e.g., "Please analyze the behavior health of diners according to the whole video").

Crucially, these tiers indicate the scale and type of perceptual aggregation required rather than an absolute ordering of difficulty: Level-3 items demand more dimensions of evidence, but models can sometimes achieve high L3 scores by producing plausible, prior-consistent text that matches the programmatic label without precise frame-level grounding. The overall distribution follows a pyramid (L1:L2:L3 $\approx$ 9:2:1), prioritizing foundational perception while keeping a targeted stress set for compositional evaluation.

4.3 Evaluation Metrics and Scoring Logic

We use a hybrid evaluation that balances objective accuracy with semantic synthesis. For L1/L2 question types (SC, TF, MC, FB) we apply exact/regex matching; for Multiple Choice we compute F1 between predicted and ground-truth option sets. L3 short answers are judged by a structured GPT-4–based evaluator (GPT-4o) that scores responses 1–10 on four axes: temporal grounding, nutritional completeness, reasoning validity, and fluency. We report per-tier accuracies and an aggregated weighted score for ranking.

5 Experiments

5.1 Experimental setup

To rigorously evaluate the state-of-the-art in eating behavior understanding, we conduct extensive experiments on EatVid-QA.

Evaluated Models. We evaluate a diverse set of video-capable large language models (Video-LLMs), organized into three categories for clarity:

- Open-source generalists: InternVL2-8B [7], Qwen2.5-VL-7B [3], LLaVA-Video-7B [25], InternVideo2.5 [58], VideoLLaMA3-7B [65] and VideoLLaMA2-7B [8]. These models were selected as representative open-source multimodal understanding systems.
- Specialized and efficient models: MiniCPM-o-4.5 (edge-optimized) and Molmo2-8B (grounding-optimized) [10, 63], chosen for their low-latency and task-focused designs.
- Proprietary API-based models (upper-bound references): Qwen3.5-35B-A3B (API) and Qwen3-VL-32B (API). We also evaluate additional non-Qwen API models, including Gemini-2.5-Flash and Seed-API, on the full EatVid-QA set; these results are reported in the Supplementary. We include proprietary systems purely as performance upper bounds.

Implementation Details. To ensure a fair comparison across Video-LLMs with diverse temporal processing capacities, we adopt a single, standardized input protocol for the main evaluation: 32 uniformly sampled frames per trial to avoid inconsistencies and potential evaluation bias.

Concretely, any session that substantially exceeds the canonical context (we exclude extreme long sessions—those longer than 5 minutes—from the primary evaluation split) is omitted from the main 32-frame benchmark to preserve comparability across models. These excluded long sessions are retained for targeted analyses reported in the Supplementary.

For models that can reasonably accept sparse, long-range inputs, we run an additional control condition in which the full session is sampled at 2 FPS across the entire video. To verify robustness to frame density, we also evaluate matched settings with uniform 32-frame and 64-frame sampling, as well as 1, 2, and 4 FPS sampling. These experiments confirm that denser sampling provides only marginal improvements and that the fundamental perception–reasoning gap persists across temporal input densities. Detailed results are reported in the Supplementary.

Evaluation protocol. Unless otherwise noted, all models are evaluated in a zero-shot setting to measure inherent domain-transfer ability. We use a single, fixed prompt template for all models (prompt templates are provided in the Supplementary). Reported metrics include: raw accuracy (Acc), difficulty-weighted accuracy (WAcc), short-answer weighted accuracy (SA-WAcc), and objective-question weighted accuracy (Non-SA-WAcc). Statistical significance testing and confidence intervals are provided in Supplementary tables.

Table 3: Evaluation of Representative Video-LLMs on the Full Dataset ($N = 6300$). **Acc** denotes raw accuracy, **WAcc** is the weighted accuracy across L1-L3. **SA-WAcc** refers to the weighted score on Short Answer (SA) questions, while **Non-SA-WAcc** covers objective types (SC, TF, MC, FB).

Model	Acc ↑	WAcc ↑	SA-WAcc ↑	Non-SA-WAcc ↑
InternVL2-8B	<u>0.3787</u>	**0.6537**	**0.8462**	<u>0.2843</u>
Qwen2.5-VL-7B	0.3310	<u>0.6300</u>	<u>0.8323</u>	0.2415
VideoLLaMA2-7B	0.3332	0.5530	0.7101	0.2516
VideoLLaMA3-7B	**0.3868**	0.4678	0.5404	**0.3284**
LLaVA-Next-Video-7B	0.3092	0.3094	0.3333	0.2633
InternVideo2.5-8B	0.2476	0.2776	0.3235	0.1895

5.2 Main results and discussion

Tabs. 3 and 4 summarize model performance on the full EatVid-QA set and on the Hard subset. The Hard subset is constructed by selecting task instances

that received the lowest aggregate scores across an initial evaluation sweep of a representative model group — i.e., tasks that are consistently challenging for multiple architectures — to form a diagnostic "stress" split for more extensive evaluation (including models that were subsequently fine-tuned).

For experiments on this Hard subset only, we increase the temporal resolution of model inputs by sampling 64 frames per input clip (uniformly sampled across each video) to better expose long-range temporal cues that distinguish these difficult instances. The main evaluation on the full set retains the original frame-sampling setting.

Additional full-set API baselines in the Supplementary show the same split between fluent synthesis and grounded perception: Qwen3.5-35B-A3B reaches 96.7 SA-WAcc but 30.6 Non-SA-WAcc, while Gemini-2.5-Flash and Seed-API exhibit similar SA–Non-SA gaps. This confirms that the observed perception–reasoning imbalance is not specific to a single model family.

Table 4: Comparison on the **Hard Subset** ($N = 376$), focusing on compositional reasoning and behavioral health assessment. Models are ranked by Weighted Accuracy (**WAcc**).

Model	Params	Acc ↑	WAcc ↑	SA-WAcc ↑	Non-SA-WAcc ↑
API-based (Proprietary)					
Qwen3.5-35B-A3B	35B	**0.4388**	**0.7359**	**0.9667**	**0.2982**
Qwen3-VL-32B	32B	0.3830	0.6988	0.9333	0.2540
Open-source (Generalist)					
Molmo2	8B	0.2686	<u>0.6008</u>	**0.8333**	0.1596
InternVL2	8B	0.2606	0.5868	<u>0.8000</u>	0.1823
Qwen2.5-VL	7B	0.2500	0.5798	<u>0.8000</u>	0.1623
VideoLLaMA2	7B	<u>0.3112</u>	0.5606	0.7333	0.2329
MiniCPM-O-4.5	8B	0.2872	0.5493	0.7333	0.2002
VideoLLaMA3	7B	0.2420	0.5417	0.7333	0.1781
LLaVA-NeXT-Video	7B	0.2899	0.3248	0.3667	<u>0.2455</u>
InternVideo2.5	8B	0.1702	0.2183	0.2667	0.1264
AGCoT-LoRA (Ours)	7B	**0.4069**	**0.6130**	0.7667	**0.3214**

Model scale and iteration explain most of the observed gap. Performance differences are primarily driven by model capacity and recency of iteration rather than access modality alone. On the Hard subset, larger/recent models (e.g., Qwen3.5-35B-A3B, WAcc = 0.7359) outperform smaller or older open-source releases; AGCoT-LoRA (7B) narrows the margin (WAcc = 0.6130) but does not eliminate it. This gap is most apparent on weighted accuracy and on tasks that require long-range temporal evidence aggregation.

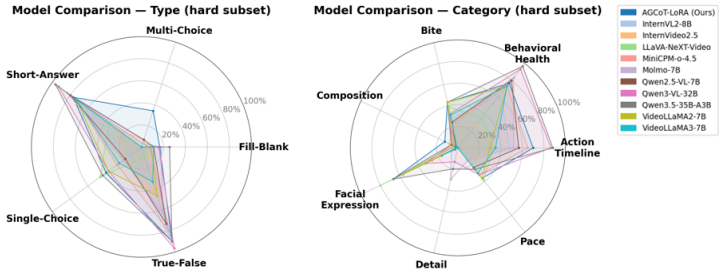

Fig. 3: Multi-dimensional performance analysis of Video-LLMs on EatVid-QA. **Left:** Performance by question type. **Right:** Performance by capability dimension.

Good perception does not imply good temporal reasoning or synthesis. High objective (non-SA) scores—indicating accurate low-level recognition—do not guarantee strong short-answer performance. For example, VideoLLaMA3-7B achieves competitive Non-SA-WAcc (0.3284) but exhibits suboptimal performance on SA-WAcc (0.5404) , implying that accurate identification of items/events alone is insufficient: models must also organize and integrate those low-level signals into temporally coherent, causal assessments.

Apparent L3 success can be driven by priors rather than evidence. High SA-WAcc is often supported by linguistic plausibility and commonsense fallback when spatiotemporal or fine-grained signals are noisy or absent. We therefore treat strong L3 scores as a hypothesis rather than proof of true evidence-based reasoning; targeted probes (e.g., evidence-masking, temporal shuffling, and frame-level grounding tests) are needed to confirm whether answers are grounded in video evidence or in learned priors.

5.3 Fine-Grained Analysis

Objective accuracy vs. linguistic fluency Fig. 3 reveals a marked gap between tasks that reward linguistic plausibility and those requiring objective precision. Models sustain relatively high scores on T/F and short-answer items ($\sim$70.6% and $\sim$80.7%, respectively), where fluent language can mask missing evidence. In contrast, accuracy drops sharply on Fill-in-the-Blank (5.8%) and Multiple-Choice (3.57%) items that demand exact counts or precise selection among close distractors. This pattern indicates limited numerical calibration and multi-object tracking in current Video-LLMs, rather than a failure of surface-level fluency.

Perception–reasoning imbalance Performance across capability dimensions is highly uneven (see right panel of Fig. 3). Models tend to score well on high-level synthesis tasks such as Behavioral Health and Action Timeline ($> 60\%$),

where semantic priors help produce plausible narratives, but perform poorly on foundational grounding tasks like Bite Detection (∼21.8%) and Dietary Composition (∼9.2%). The results point to a gap: accurate low-level action grounding is often lacking even when models produce convincing high-level interpretations.

5.4 AGCoT-LoRA: A Domain-Adapted Baseline

To bridge the perception-reasoning gap, we propose AGCoT-LoRA, which explicitly trains the model to generate a Chain-of-Thought (CoT) anchored in our Tier 1 and Tier 2 annotations.

Fine-tuning the Qwen2.5-VL-7B model on a subset of our annotated data yields substantial improvements. As shown in Tab. 4, AGCoT-LoRA achieves a raw accuracy of 0.4069, a significant leap from the zero-shot base model's 0.2500. More importantly, its Non-SA Weighted Accuracy increases from 0.1623 to 0.3214, nearly doubling the model's ability to handle objective, evidence-based questions. By forcing the model to first output reasoning details (e.g., identifying specific bite timestamps and calculating intervals), AGCoT-LoRA mitigates "hallucinatory reasoning." It shifts the model's behavior from making "lucky guesses" based on linguistic priors to making "calculated inferences" based on visual evidence. Detailed training hyperparameters, and illustrative CoT visualizations are provided in the supplementary material.

To isolate the effect of Chain-of-Thought supervision, we conduct an ablation against a plain LoRA baseline trained on the same data but without annotation-guided CoT targets. On the matched Hard subset under 4 FPS sparse sampling, AGCoT-LoRA achieves **36.4%** Non-SA-WAcc, substantially outperforming both the base Qwen2.5-VL (14.1%) and plain LoRA (22.3%). This indicates that the gains are not merely due to parameter-efficient domain adaptation, but also to explicit reasoning supervision grounded in annotation traces.

6 Discussion & Analysis

6.1 Why current Video-LLMs fall short

Our experiments reveal three dominant failure modes:

Micro-Action Localization Deficits. Current Video-LLMs often fail to recognize and precisely localize very brief, repetitive micro-actions especially under occlusion and viewpoint variation; this reveals a core weakness in short-timescale pattern recognition and event-boundary detection.

Domain-knowledge deficit. Eating behavior requires interpreting subtle, domain-specific cues (mastication rhythm, pacing, distraction patterns) that are rarely labeled in large, general-purpose video corpora; consequently, models pre-trained on broad video-text corpora lack the task-specific priors to map low-level signals to behavioral-health concepts.

Calibration under distribution shift. In specialized forced-choice or diagnostic tasks, models often exhibit option bias or overconfident predictions when visual evidence is weak or ambiguous, revealing poor calibration under domain shift.

6.2 What makes EatVid-Bench especially challenging

EatVid-Bench is intentionally anti-shortcut: correct responses frequently require cross-frame temporal counting and multimodal integration. The benchmark therefore shifts emphasis from "what is happening?" toward "how (temporal dynamics), how much (counts/pace), and what does it imply for behavioral health?" — tasks that expose weaknesses in temporal grounding, multimodal fusion, and evidence-based reasoning.

6.3 Limitations and caveats

Geographic and cultural coverage. The current release does not yet exhaustively represent global eating traditions or all cuisine types; expanding cultural and culinary diversity is a priority for future versions.

Scope of behavioral claims. Our benchmark targets behavioral and affective inference rather than precise nutritional quantification; accurate calorie or macro-gram estimation remains out of scope due to depth, occlusion, and sensor limitations.

Tier-3 semantic annotations and mitigation of LLM biases. Tier-3 labels are produced by high-capacity multimodal language models, annotation outputs undergo automated rule-based validation and targeted human review to reduce spurious or fluently phrased but unsupported statements. Nonetheless, some residual model-inherited biases and fluency-over-accuracy tendencies may persist and should be considered when interpreting downstream results.

7 Conclusion

We introduce EatVid-Bench, a large-scale video-language benchmark for behavioral health. Leveraging a novel three-tier automated pipeline, we transform around 3,000 minutes of real-world eating sessions into a structured knowledge base with 12 complementary signal dimensions. Our evaluation reveals a significant "perception–reasoning gap" in current Video-LLMs, particularly in subsecond grounding and temporal counting. We address this via AGCoT-LoRA, demonstrating that annotation-guided fine-tuning enhances the grounding of health assessments in physical signals. By providing provenance-linked annotations and standardized protocols, EatVid-Bench reorients video-language research from shortcut-based recognition toward evidence-based behavioral understanding. We expect these resources to accelerate progress in digital-health applications that require trustworthy, long-form interpretation of human behavior.

References

1. Alshurafa, N., Zhang, S., Romano, C., Zhang, H., Pfammatter, A.F., Lin, A.W.: Association of number of bites and eating speed with energy intake: Wearable technology results under free-living conditions. Appetite **167**, 105653 (2021)
2. Asil, E., Erkmen, E.: Nutrition is associated with violent and criminal behaviors. Current Nutrition Reports **14**(1), 76 (2025)
3. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., Zhong, H., Zhu, Y., Yang, M., Li, Z., Wan, J., Wang, P., Ding, W., Fu, Z., Xu, Y., Ye, J., Zhang, X., Xie, T., Cheng, Z., Zhang, H., Yang, Z., Xu, H., Lin, J.: Qwen2.5-vl technical report (2025), https://arxiv.org/abs/2502.13923
4. Bell, B.M., Alam, R., Alshurafa, N., Thomaz, E., Mondol, A.S., de la Haye, K., Stankovic, J.A., Lach, J., Spruijt-Metz, D.: Automatic, wearable-based, in-field eating detection approaches for public health research: a scoping review. NPJ digital medicine **3**(1), 38 (2020)
5. Chen, G., Liu, Y., Huang, Y., He, Y., Pei, B., Xu, J., Wang, Y., Lu, T., Wang, L.: Cg-bench: Clue-grounded question answering benchmark for long video understanding. arXiv preprint arXiv:2412.12075 (2024)
6. Chen, L.H., Lu, S., Zeng, A., Zhang, H., Wang, B., Zhang, R., Zhang, L.: Motionllm: Understanding human behaviors from human motions and videos. IEEE Transactions on Pattern Analysis and Machine Intelligence (2025)
7. Chen, Z., Wang, W., Cao, Y., Liu, Y., Gao, Z., Cui, E., Zhu, J., Ye, S., Tian, H., Liu, Z., et al.: Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. arXiv preprint arXiv:2412.05271 (2024)
8. Cheng, Z., Leng, S., Zhang, H., Xin, Y., Li, X., Chen, G., Zhu, Y., Zhang, W., Luo, Z., Zhao, D., et al.: Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms. arXiv preprint arXiv:2406.07476 (2024)
9. Chung, J.L., Ong, L.Y., Leow, M.: Comparative analysis of skeleton-based human pose estimation. Future Internet **14**, 380 (2022). https://doi.org/10.3390/fi14120380
10. Clark, C., Zhang, J., Ma, Z., Park, J.S., Salehi, M., Tripathi, R., Lee, S., Ren, Z., Kim, C.D., Yang, Y., et al.: Molmo2: Open weights and data for vision-language models with video understanding and grounding. arXiv preprint arXiv:2601.10611 (2026)
11. Dakanalis, A., Mentzelou, M., Papadopoulou, S.K., Papandreou, D., Spanoudaki, M., Vasios, G.K., Pavlidou, E., Mantzorou, M., Giaginis, C.: The association of emotional eating with overweight/obesity, depression, anxiety/stress, and dietary patterns: a review of the current clinical evidence. Nutrients **15**(5), 1173 (2023)
12. Damen, D., Doughty, H., Farinella, G., Fidler, S., Furnari, A., Kazakos, E., Moltisanti, D., Munro, J., Perrett, T., Price, W., Wray, M.: Scaling egocentric vision: The epic-kitchens dataset. ArXiv **abs/1804.02748** (2018)
13. Damen, D., Doughty, H., Farinella, G., Fidler, S., Furnari, A., Kazakos, E., Moltisanti, D., Munro, J., Perrett, T., Price, W., Wray, M.: The epic-kitchens dataset: Collection, challenges and baselines. IEEE Transactions on Pattern Analysis and Machine Intelligence **43**, 4125–4141 (2020). https://doi.org/10.1109/tpami.2020.2991965
14. Ekici, E.M., Mengi Çelik, Ö., Metin, Z.E.: The relationship between night eating behavior, gastrointestinal symptoms, and psychological well-being: insights from a cross-sectional study in türkiye. Journal of Eating Disorders **13**(1), 14 (2025)

15. Fu, C., Dai, Y., Luo, Y., Li, L., Ren, S., Zhang, R., Wang, Z., Zhou, C., Shen, Y., Zhang, M., et al.: Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 24108–24118 (2025)
16. Grauman, K., Westbury, A., Byrne, E., Chavis, Z., Furnari, A., Girdhar, R., Hamburger, J., Jiang, H., Liu, M., Liu, X., et al.: Ego4d: Around the world in 3,000 hours of egocentric video. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 18995–19012 (2022)
17. Hamurcu, P., Köse, G., Şahin, İ.N., Çelik, A.: Dört faktörlü yeme farkındalığı ölçeği'nin türkçeye uyarlanması: geçerlik ve güvenirlik çalışması. Türkiye Klinikleri J Health Sci **8**(2), 187–98 (2023)
18. Heidari, M., Khodadadi Jokar, Y., Madani, S., Shahi, S., Shahi, M.S., Goli, M.: Influence of food type on human psychological–behavioral responses and crime reduction. Nutrients **15**(17), 3715 (2023)
19. Jiang, L., Qiu, B., Liu, X., Huang, C., Lin, K.: Deepfood: food image analysis and dietary assessment via deep model. IEEE Access **8**, 47477–47489 (2020)
20. Kesen, I., Pedrotti, A., Dogan, M., Cafagna, M., Acikgoz, E.C., Parcalabescu, L., Calixto, I., Frank, A., Gatt, A., Erdem, A., et al.: Vilma: A zero-shot benchmark for linguistic and temporal grounding in video-language models. arXiv preprint arXiv:2311.07022 (2023)
21. Korb, S., Massaccesi, C., Gartus, A., Lundström, J.N., Rumiati, R., Eisenegger, C., Silani, G.: Facial responses of adult humans during the anticipation and consumption of touch and food rewards. Cognition **194**, 104044 (2020)
22. Kyritsis, K., Diou, C., Delopoulos, A.: A data driven end-to-end approach for in-the-wild monitoring of eating behavior using smartwatches. IEEE Journal of Biomedical and Health Informatics **25**(1), 22–34 (2020)
23. Kyritsis, K.A., Diou, C., Delopoulos, A.: Modeling wrist micromovements to measure in-meal eating behavior from inertial sensor data. IEEE Journal of Biomedical and Health Informatics **23**, 2325–2334 (2019). https://doi.org/10.1109/jbhi.2019.2892011
24. Li, K., Wang, Y., He, Y., Li, Y., Wang, Y., Liu, Y., Wang, Z., Xu, J., Chen, G., Luo, P., et al.: Mvbench: A comprehensive multi-modal video understanding benchmark. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22195–22206 (2024)
25. Lin, B., Ye, Y., Zhu, B., Cui, J., Ning, M., Jin, P., Yuan, L.: Video-llava: Learning united visual representation by alignment before projection. In: Proceedings of the 2024 conference on empirical methods in natural language processing. pp. 5971–5984 (2024)
26. Lin, K.Q., Zhang, P., Chen, J., Pramanick, S., Gao, D., Wang, A.J., Yan, R., Shou, M.Z.: Univtg: Towards unified video-language temporal grounding. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 2794–2804 (2023)
27. Linseisen, J., Renner, B., Gedrich, K., Wirsam, J., Holzapfel, C., Lorkowski, S., Watzl, B., Daniel, H., Leitzmann, M., et al.: Data in personalized nutrition: Bridging biomedical, psycho-behavioral, and food environment approaches for population-wide impact. Advances in Nutrition **16**(7), 100377 (2025)
28. Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Jiang, Q., Li, C., Yang, J., Su, H., et al.: Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In: European conference on computer vision. pp. 38–55. Springer (2024)

29. Liu, Y., Ma, Z., Qi, Z., Wu, Y., Shan, Y., Chen, C.W.: E.t. bench: Towards open-ended event-level video-language understanding. ArXiv **abs/2409.18111** (2024). https://doi.org/10.48550/arxiv.2409.18111
30. Lugaresi, C., Tang, J., Nash, H., McClanahan, C., Uboweja, E., Hays, M., Zhang, F., Chang, C.L., Yong, M.G., Lee, J., et al.: Mediapipe: A framework for building perception pipelines. arXiv preprint arXiv:1906.08172 (2019)
31. Maaz, M., Rasheed, H., Khan, S., Khan, F.: Video-chatgpt: Towards detailed video understanding via large vision and language models. In: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 12585–12602 (2024)
32. Mao, R., He, J., Lin, L., Shao, Z., Eicher-Miller, H.A., Zhu, F.: Improving dietary assessment via integrated hierarchy food classification. In: 2021 IEEE 23rd International Workshop on Multimedia Signal Processing (MMSP). pp. 1–6. IEEE (2021)
33. Maqsood, S., Ahmed, F., Arshad, M.T., Ikram, A., Abdullahi, M.A.: Comparative analysis of food addiction and obesity: A critical review. Food Science & Nutrition **13**(8), e70799 (2025)
34. Masram, P.N.: Assess the effectiveness of self-instructional module on knowledge regarding effects of junk food on mental health among the undergraduate students at selected colleges. International Journal of Science and Research (IJSR) (2023), https://api.semanticscholar.org/CorpusID:264938936
35. McClelland, J., Dalton, B., Kekic, M., Bartholdy, S., Campbell, I.C., Schmidt, U.: A systematic review of temporal discounting in eating disorders and obesity: Behavioural and neuroimaging findings. Neuroscience & Biobehavioral Reviews **71**, 506–528 (2016)
36. Melo, G.L.R., Santo, R.E., Clavel, E.M., Prous, M.B., Koehler, K., Vidal-Alaball, J., van der Waerden, J., Gobiņa, I., López-Gil, J.F., Lima, R., et al.: Digital dietary interventions for healthy adolescents: A systematic review of behavior change techniques, engagement strategies, and adherence. Clinical nutrition **45**, 176–192 (2025)
37. Nakamura, A., Saito, T., Ikeda, D., Ohta, K., Mineno, H., Nishimura, M.: Automatic detection of chewing and swallowing †. Sensors (Basel, Switzerland) **21** (2021). https://doi.org/10.3390/s21103378
38. Ning, M., Zhu, B., Xie, Y., Lin, B., Cui, J., Yuan, L., Chen, D., Yuan, L.: Video-bench: A comprehensive benchmark and toolkit for evaluating video-based large language models. Computational Visual Media (2025)
39. O'Hara, C., Gibney, E.R.: Meal pattern analysis in nutritional science: recent methods and findings. Advances in nutrition **12**(4), 1365–1378 (2021)
40. Plizzari, C., Tonioni, A., Xian, Y., Kulshrestha, A., Tombari, F.: Omnia de egotempo: Benchmarking temporal understanding of multi-modal llms in egocentric videos. 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 24129–24138 (2025). https://doi.org/10.1109/cvpr52734.2025.02247
41. Radford, A., Kim, J.W., Xu, T., Brockman, G., McLeavey, C., Sutskever, I.: Robust speech recognition via large-scale weak supervision. In: International conference on machine learning. pp. 28492–28518. PMLR (2023)
42. Ranftl, R., Lasinger, K., Hafner, D., Schindler, K., Koltun, V.: Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. IEEE transactions on pattern analysis and machine intelligence **44**(3), 1623–1637 (2020)

43. Ravi, N., Gabeur, V., Hu, Y.T., Hu, R., Ryali, C.K., Ma, T., Khedr, H., Rädle, R., Rolland, C., Gustafson, L., Mintun, E., Pan, J., Alwala, K.V., Carion, N., Wu, C.Y., Girshick, R.B., Doll'ar, P., Feichtenhofer, C.: Sam 2: Segment anything in images and videos. ArXiv **abs/2408.00714** (2024). https://doi.org/10.48550/arxiv.2408.00714
44. Raza, M.A., Chen, L., Nanbo, L., Fisher, R.B.: Eatsense: Human centric, action recognition and localization dataset for understanding eating behaviors and quality of motion assessment. Image and Vision Computing **137**, 104762 (2023)
45. Rouast, P.V., Heydarian, H., Adam, M., Rollo, M.: Oreba: A dataset for objectively recognizing eating behavior and associated intake. IEEE Access **8**, 181955–181963 (2020). https://doi.org/10.1109/access.2020.3026965
46. Shah, M.W., Yan, Q., Pan, D., Sun, G.: Psychometric evaluation of eating behaviors and mental health among university students in china and pakistan: A cross-cultural study. Nutrients **17**(5), 795 (2025)
47. Shen, Y., Salley, J., Muth, E., Hoover, A.: Assessing the accuracy of a wrist motion tracking method for counting bites across demographic and food variables. IEEE Journal of Biomedical and Health Informatics **21**, 599–606 (2017). https://doi.org/10.1109/jbhi.2016.2612580
48. Stover, P.J., Field, M.S., Andermann, M.L., Bailey, R.L., Batterham, R.L., Cauffman, E., Frühbeck, G., Iversen, P.O., Starke-Reed, P., Sternson, S.M., et al.: Neurobiology of eating behavior, nutrition, and health. Journal of Internal Medicine **294**(5), 582–604 (2023)
49. Talekar, C.A., Chiruvella, S.: A study of dietary habits and mental health wellbeing in young adults. International Journal For Multidisciplinary Research (2023), https://api.semanticscholar.org/CorpusID:259604139
50. Tang, Y., Bi, J., Xu, S., Song, L., Liang, S., Wang, T., Zhang, D., An, J., Lin, J., Zhu, R., et al.: Video understanding with large language models: A survey. IEEE Transactions on Circuits and Systems for Video Technology (2025)
51. Team, O.: Internvl2: Better than the best—expanding performance boundaries of open-source multimodal models with the progressive scaling strategy (2024)
52. Tufano, M., Lasschuijt, M., Chauhan, A., Feskens, E., Camps, G.: Capturing eating behavior from video analysis: A systematic review. Nutrients **14** (2022). https://doi.org/10.3390/nu14224847
53. Tufano, M., Lasschuijt, M., Chauhan, A., Feskens, E.J.M., Camps, G.: Rule-based systems to automatically count bites from meal videos. Frontiers in Nutrition **11** (2024). https://doi.org/10.3389/fnut.2024.1343868
54. Vondrick, C., Patterson, D.J., Ramanan, D.: Efficiently scaling up crowdsourced video annotation. International Journal of Computer Vision **101**, 184 – 204 (2012). https://doi.org/10.1007/s11263-012-0564-1
55. Wang, C., Kumar, T.S., De Raedt, W., Camps, G., Hallez, H., Vanrumste, B.: Eat-radar: Continuous fine-grained intake gesture detection using fmcw radar and 3d temporal convolutional network with attention. IEEE Journal of Biomedical and Health Informatics **28**(2), 1000–1011 (2023)
56. Wang, P., Bai, S., Tan, S., Wang, S., Fan, Z., Bai, J., Chen, K., Liu, X., Wang, J., Ge, W., et al.: Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. arXiv preprint arXiv:2409.12191 (2024)
57. Wang, X., Zhang, Y., Zohar, O., Yeung-Levy, S.: Videoagent: Long-form video understanding with large language model as agent. In: European Conference on Computer Vision. pp. 58–76. Springer (2024)

58. Wang, Y., Li, X., Yan, Z., He, Y., Yu, J., Zeng, X., Wang, C., Ma, C., Huang, H., Gao, J., et al.: Internvideo2. 5: Empowering video mllms with long and rich context modeling. arXiv preprint arXiv:2501.12386 (2025)
59. Weng, Y., Han, M., He, H., Chang, X., Zhuang, B.: Longvlm: Efficient long video understanding via large language models. In: European Conference on Computer Vision. pp. 453–470. Springer (2024)
60. Wu, H., Li, D., Chen, B., Li, J.: Longvideobench: A benchmark for long-context interleaved video-language understanding. Advances in Neural Information Processing Systems **37**, 28828–28857 (2024)
61. Wu, J., Liu, W., Liu, Y., Liu, M., Nie, L., Lin, Z., Chen, C.W.: A survey on video temporal grounding with multimodal large language model. IEEE Transactions on Pattern Analysis and Machine Intelligence (2025)
62. Wu, J., Liu, W., Liu, Y., Liu, M., Nie, L., Lin, Z., Chen, C.W.: A survey on video temporal grounding with multimodal large language model. IEEE Transactions on Pattern Analysis and Machine Intelligence **48**, 1521–1541 (2025). https://doi.org/10.1109/tpami.2025.3615586
63. Yao, Y., Yu, T., Zhang, A., Wang, C., Cui, J., Zhu, H., Cai, T., Li, H., Zhao, W., He, Z., et al.: Minicpm-v: A gpt-4v level mllm on your phone. arXiv preprint arXiv:2408.01800 (2024)
64. Yildirim, E., Akbulut, F.P., Catal, C.: Analysis of facial emotion expression in eating occasions using deep learning. Multimedia Tools and Applications **82**(20), 31659–31671 (2023)
65. Zhang, B., Li, K., Cheng, Z., Hu, Z., Yuan, Y., Chen, G., Leng, S., Jiang, Y., Zhang, H., Li, X., et al.: Videollama 3: Frontier multimodal foundation models for image and video understanding. arXiv preprint arXiv:2501.13106 (2025)
66. Zhang, Y., Wu, J., Li, W., Li, B., Ma, Z., Liu, Z., Li, C.: Llava-video: Video instruction tuning with synthetic data. arXiv preprint arXiv:2410.02713 (2024)
67. Zhao, J., Yang, Q., Peng, Y.X., Bai, D., Yao, S., Sun, B., Chen, X., Fu, S., chen, W., Wei, X., Bo, L.: Humanomni: A large vision-speech language model for human-centric video understanding. ArXiv **abs/2501.15111** (2025). https://doi.org/10.48550/arxiv.2501.15111
68. Zhou, J., Shu, Y., Zhao, B., Wu, B., Xiao, S., Yang, X., Xiong, Y., Zhang, B., Huang, T., Liu, Z.: Mlvu: A comprehensive benchmark for multi-task long video understanding. arXiv preprint arXiv:2406.04264 **2**(5), 6 (2024)
69. Zhou, T., Chen, D., Jiao, Q., Ding, B., Li, Y., Shen, Y.: Humanvbench: Exploring human-centric video understanding capabilities of mllms with synthetic benchmark data. ArXiv **abs/2412.17574** (2024). https://doi.org/10.48550/arxiv.2412.17574