ETHOBEHAVIOR: LEARNING COMPOSITIONAL ETHOGRAM FOR PRIMATE BEHAVIOR UNDERSTANDING

Anonymous authors
Paper under double-blind review

ABSTRACT

Primate behavior understanding seeks to identify and temporally characterize behavioral events and states from video, providing scalable measurements for research in cognition and sociality. Existing methods operate within independent ethograms whose categories differ in granularity and annotation scope, limiting the comparability of behavioral measurements across studies. Although Video-Language Models (VLMs) capture fine-grained visual cues, they typically treat ethogram categories as atomic concepts and remain agnostic to the structured behavioral semantics. To address these limitations, we introduce Compositional Ethogram Factorization, a shared representation that models primate behaviors as structured combinations of reusable behavioral predicates and typed attributes. We further instantiate it in PrInstruct-34K, a multi-task video instruction dataset comprising 34K examples from four primate behavior benchmarks. It retains 58 ethogram categories and standardizes their operational definitions into three complementary tasks: multiple-choice question answering, video captioning, and multi-label temporal parsing. Building on this, we develop EthoBehavior, a VLM with an Ethogram Composition Adapter that estimates factor distributions from video under partial supervision and encodes them as continuous Ethogram Factor Tokens for free-form behavior description and structured behavior analysis. Experiments across all three tasks and on unseen factor compositions show that EthoBehavior consistently outperforms instruction-tuned baselines and large open-source VLMs including GLM-5.3-Flash and DeepSeek-V4.1-Flash. Together, these results establish compositional ethogram representations as a principled basis for generalizable and comparable primate behavior analysis.

1 INTRODUCTION

Quantitative analysis of non-human primate behavior underpins research in cognition, social organization, and behavioral evolution and animal conservation. Recent video understanding models Ma et al. (2023); Brookes et al. (2024a) have expanded automated primate behavior analysis from category recognition to behavior description and temporal characterization across large observational archives. Current primate video benchmarks Ma et al. (2023); Brookes et al. (2024b); Duporge et al. (2025) formulate this task through distinct ethograms, whose categories serve as supervision and targets. Because these ethograms were developed for different scientific questions and observational protocols, they encode behavior at different levels of abstraction and specify different semantics. For example, one ethogram may encode *locomotion* as a higher-level behavioral category, whereas another distinguishes specific motor patterns such as *walking or jumping*. Consequently, semantic correspondences across benchmarks are rarely one-to-one: direct label matching misses partial correspondences, whereas collapsing categories into a common label space obscures distinctions retained by individual ethograms. This raises a fundamental representation problem:

> *How can heterogeneous ethograms be represented in a shared behavioral space while preserving the ethogram-specific semantics of their categories?*

Conventional behavior recognizers Yan et al. (2018); Ma et al. (2023); Liu et al. (2026) learn specific functions over fixed category and therefore inherit the semantic boundaries of their ethograms.

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

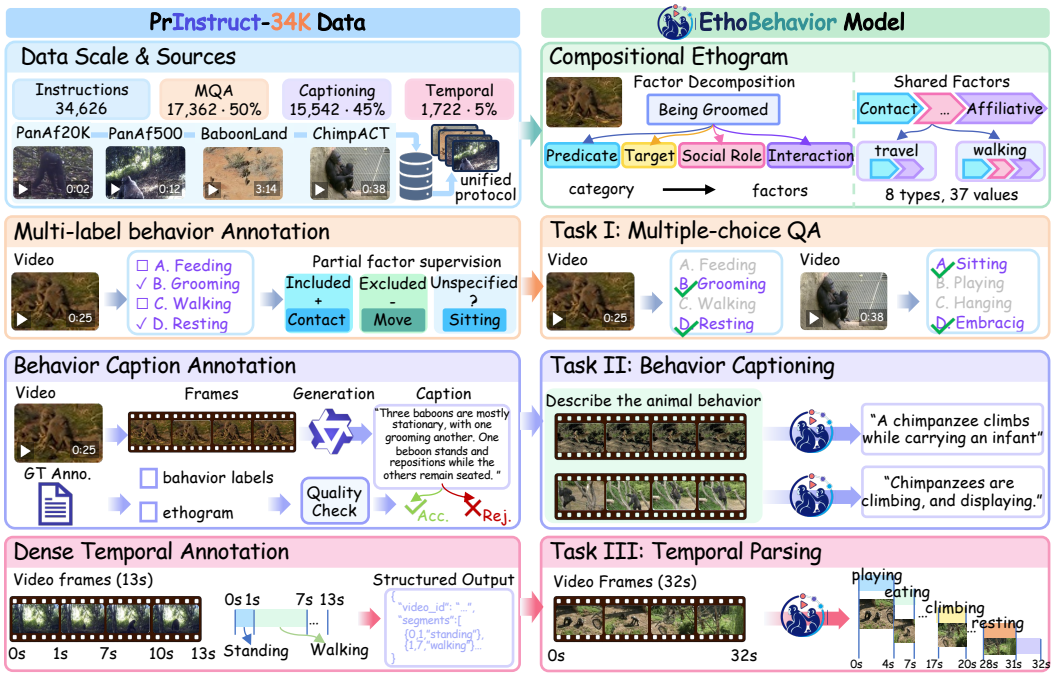


Figure 1: Overview of PrInstruct-34K and EthoBehavior. The left panel shows how annotations from four primate benchmarks support behavior question answering, captioning, and temporal parsing. The right panel illustrates how ethogram-specific categories connect to shared behavioral factors that EthoBehavior uses across these tasks.

Ethogram-aware models Brookes et al. (2024a); Jing et al. (2026) enrich these categories with textual cues or align them to canonical ontologies with hierarchical relations. However, the complete behavior category remains the basis of supervision and prediction, such that semantic dimensions shared by non-equivalent categories are not exposed as independently reusable units. In parallel, recent VLMs Qi et al. (2026); Clark et al. (2026) demonstrate detailed video description and fine-grained reasoning. These capabilities do not by themselves establish a consistent mapping between model outputs and the semantic distinctions encoded by different ethograms. Consequently, existing formulations lack an explicit representation that captures behavior shared ethograms while retaining the individual semantic distinctions.

Our key insight is that the semantics of ethograms can be expressed compositionally through recurring, typed behavioral factors. For example, *walking* and *climbing up* share the predicate *move* but differ in locomotion mode and direction. Accordingly, we represent each ethogram category using the subset of a shared factor vocabulary supported by its definition. The original category spaces are retained for annotation and evaluation, while factors not determined by a category definition remain unspecified. This enables factor level semantic sharing across ethograms while preserving distinctions among the original categories.

We introduce Compositional Ethogram Factorization (CEF), a structured representation that decomposes the definition of each ethogram category into a partial assignment over reusable behavioral factors. To support learning and evaluation across complementary ethogram behavior understanding, we construct PrInstruct-34K with three tasks: multiple-choice question answering, behavior captioning and dense temporal parsing. The captions are generated by Qwen3.5-397B Team (2026), while another targets are derived from the original annotations. We instantiate CEF in EthoBehavior, an VLM equipped with an Ethogram Composition Adapter (ECA) that estimates factor distributions from visual representation. ECA subsequently encodes the predicted factor distributions as Ethogram Factor Tokens (EFTs), which augment the original visual tokens and condition the language model on explicit ethogram structure. The resulting representation supports both natural-language behavior description and schema-constrained behavioral analysis. We evaluate EthoBehavior across all three PrInstruct-34K tasks and on unseen factor compositions whose constituent values are also excluded from task training.

In summary, our contributions are as follows:

- We introduce **Compositional Ethogram Factorization (CEF)**, which represents heterogeneous ethogram categories as partial assignments over a shared vocabulary of reusable, typed behavioral factors, enabling cross-ethogram semantic sharing while preserving ethogram-specific distinctions.
- We construct **PrInstruct-34K**, a multi-task video instruction dataset of 34K examples that retains 58 categories from four primate behavior benchmarks and standardizes their operational definitions across multiple-choice question answering, behavior captioning, and multi-label temporal parsing.
- We develop **EthoBehavior**, a VLM that uses an Ethogram Composition Adapter to infer factor distributions and encode them as Ethogram Factor Tokens. Experiments across all three tasks and unseen action compositions demonstrate consistent improvements over instruction-tuned baselines and large open-source VLMs.

2 Related Works

2.1 Animal Behavior Understanding

Markerless pose estimation and tracking recover body configurations, trajectories, and identities Mathis et al. (2018); Pereira et al. (2022), while supervised classifiers and unsupervised discovery map them to expert labels or recurring motifs Bohnslav et al. (2021); Weinreb et al. (2024). These methods scale behavioral measurement but remain sensitive to species, recording conditions, and label granularity, limiting cross-dataset reuse. Primate interactions require identity, role, and temporal context Ma et al. (2023), while wild footage introduces occlusion, vegetation, and rare behaviors Brookes et al. (2024b); Vogg et al. (2025). These distinctions matter for studies of primate sociality, communication, cognition, and welfare because different labels can change which interactions are measured. Benchmarks thus span controlled and wild settings with heterogeneous labels Bala et al. (2020); Yao et al. (2023); Bain et al. (2021), motivating shared representations that preserve ethogram-specific distinctions.

2.2 Structured Representations of Behavioral Ethograms

Classical ethograms use expert-defined codes and sampling schemes, although the boundaries and granularity of behavior remain debated Levitis et al. (2009); Garcia-Loya et al. (2026). Top-down classifiers scale expert labels to frames Kabra et al. (2013); Bohnslav et al. (2021), whereas bottom-up methods recover motion motifs and their temporal organization Berman et al. (2014); Wiltschko et al. (2015); Hsu & Yttri (2021); Luxem et al. (2022); Weinreb et al. (2024). Ontologies standardize behavior classes and phenotype relations Gkoutos et al. (2012; 2015). Yet classifiers remain tied to dataset labels, motifs lack stable operational semantics, and ontologies generally align complete categories. Taken together, these limitations motivate a compositional ethogram representation that reuses factors for behavior understanding.

3 From Ethogram Categories to Compositional Factors

We propose Compositional Ethogram Factorization (CEF) to represent specific behavior categories with shared behavioral factors. A category is a label defined by an individual ethogram and assigned to a video clip, whereas a factor is a reusable attribute with a fixed vocabulary. CEF retains the original categories as prediction targets while sharing factor supervision across ethograms.

Factor definitions. The factors are derived from the annotation guidelines Brookes et al. (2024b); Ma et al. (2023); Duporge et al. (2025) and organized using predicate-argument and compositional action representations Yatskar et al. (2016); Sadhu et al. (2021); Schenk et al. (2025); Materzynska et al. (2020). CEF summarizes eight factors as follows: `predicate` records the core action or state, such as *move*, *feed*, *groom*, or *rest*. `locomotion_mode` records the motor pattern used for locomotion, such as *walk*, *run*, or *climb*. `direction` records an explicitly labeled movement

direction. `posture` records the focal individual's gross body configuration. `target` identifies the entity toward which a behavior is directed, while `social_role` distinguishes asymmetric roles such as actor and recipient. Finally, `interaction_mode` records the form of a social interaction, and `signal_mode` records the form of communication.

The mapping is category-specific but reuses values across ethograms. In the example, *locomotion* is annotated with *move* and leaves the remaining factors unknown. *Walking* adds *walk*, while *climbing up* adds *climb* and *upward*. CEF therefore records their shared information without merging the three original labels.

Factor spaces. Let $\mathbb{D} = \{1, \ldots, N_{\text{eth}}\}$ index the ethograms, and let $\mathbb{C}_d$ denote the category set defined by ethogram $d \in \mathbb{D}$. We define the joint category space as

$$\mathbb{C} = \{c = (d, \ell) \mid d \in \mathbb{D},\ \ell \in \mathbb{C}_d\}, \tag{1}$$

where $c = (d, \ell)$ pairs a category label with its ethogram identifier. This pairing prevents labels with the same name but different annotation rules from being treated as identical.

Let $\mathbb{K} = \{1, \ldots, 8\}$ index the eight factors. Factor k has a pre-defined vocabulary $\mathbb{V}_k = \{v_{k,1}, \ldots, v_{k,n_k}\}$. Let $\mathbb{U} = \{(k, j) \mid k \in \mathbb{K},\ 1 \le j \le n_k\}$ index all factor values. CEF maps each category to one or more clauses, where a clause lists the factor values required by one interpretation of the category definition:

$$\phi(c) = \{A_{c,1}, \ldots, A_{c,r_c}\}, \qquad A_{c,h} \subseteq \mathbb{U}, \quad r_c \ge 1. \tag{2}$$

Most categories have one clause; a definition with alternatives has one clause per alternative. We distinguish values entailed by every alternative from values mentioned by at least one alternative:

$$\text{Ent}(c) = \bigcap_{A \in \phi(c)} A, \qquad \text{Poss}(c) = \bigcup_{A \in \phi(c)} A. \tag{3}$$

Factors omitted from a clause are unspecified; omission does not assert their absence. For example, a category defined as *walking or running* entails *move*, while *walk* and *run* are only possible values.

Partial factor supervision. For an annotated video instance i from ethogram d_i, let $\mathbb{P}_i^{\text{cat}} \subseteq \{(d_i, \ell) : \ell \in \mathbb{C}_{d_i}\}$ be its positive categories, pooled across annotated intervals when necessary. The entailed factor values are $\mathbb{F}_i = \bigcup_{c \in \mathbb{P}_i^{\text{cat}}} \text{Ent}(c)$. We infer a negative value only when a full category vector marks every category that could express it as negative; let $\mathbb{N}_i$ contain such values. This rule assumes the relevant individuals and intervals were annotated. Values specific to one alternative or unsupported by complete annotation remain unknown. For $u = (k, j)$, the target and observation mask are

$$y_{i,j}^{(k)} = \mathbb{I}[u \in \mathbb{F}_i], \qquad m_{i,j}^{(k)} = \mathbb{I}[u \in \mathbb{F}_i \cup \mathbb{N}_i]. \tag{4}$$

Thus $(y, m) = (1, 1)$ denotes an entailed value, $(0, 1)$ a confirmed negative, and $(0, 0)$ an unspecified value. Appendix A specifies the negative-inference rule.

4 PRINSTRUCT-34K

Building on these, we construct **PrInstruct-34K**, a video-instruction dataset that converts the behavioral annotations of PanAf20K and PanAf500 Brookes et al. (2024b), BaboonLand Duporge et al. (2025), and ChimpACT Ma et al. (2023) into a common instruction format. The dataset contains 34,626 instruction instances constructed from 17,362 qualified video records. Because one video may support multiple task formulations, the number of instances does not represent the number of unique videos. PrInstruct-34K retains the 58 categories and their definitions rather than merging them into a universal class inventory. It supports multiple-choice question answering, behavior captioning, and dense temporal parsing. Details are given in Appendix B (Table 4).

4.1 TASK FORMULATION

We formulate PrInstruct-34K around three tasks to evaluate complementary capabilities rather than treating behavior understanding as a single classification problem. The task design varies two dimensions: its semantic form and its temporal resolution. Multiple-choice question answering evaluates categorical identification within an ethogram-specific category set, answering *what behaviors*

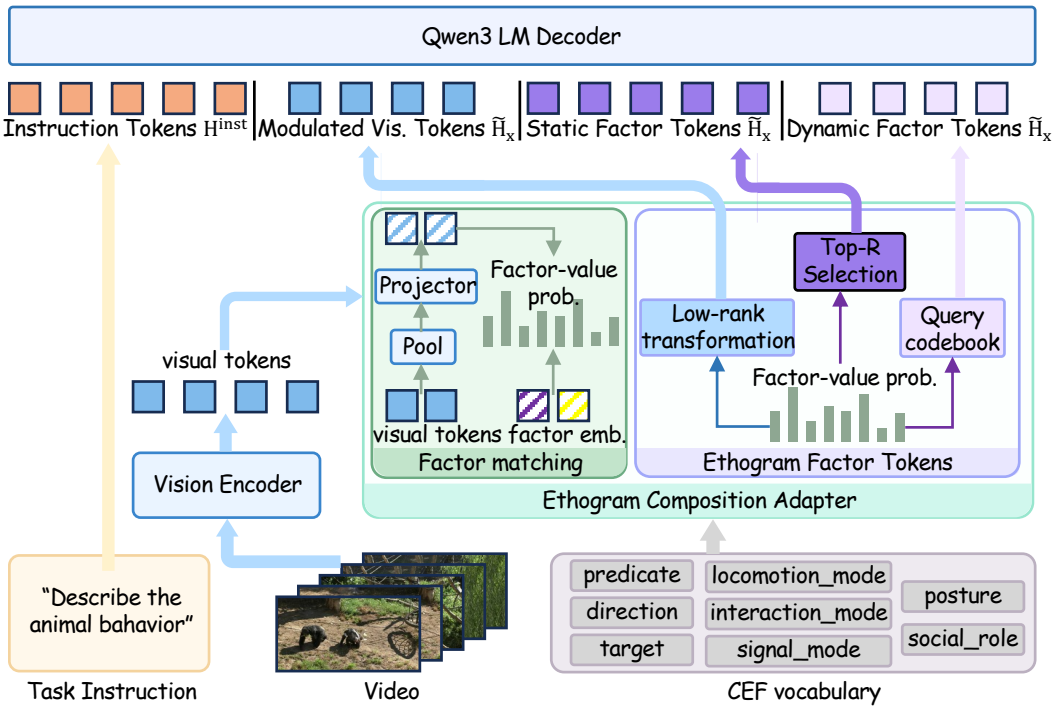


Figure 2: Framework of our EthoBehavior. Video features and task instruction enter the Ethogram Composition Adapter (ECA), which estimates values in the shared compositional ethogram. Named factor values form static Ethogram Factor Tokens (EFTs), while a residual codebook yields dynamic EFTs. Both token types join the factor-modulated visual tokens and instruction tokens for multiple-choice question answering, captioning, and temporal parsing.

occur. Behavior captioning replaces the finite category output with a natural-language description, answering *how is behavior described*. Temporal parsing further increases the temporal resolution from the complete video to individual intervals, answering *when and how do behaviors evolve over time*. Together, the three tasks evaluate categorical recognition, linguistic articulation, and temporal organization of behavioral observations.

4.2 Dataset Construction & Quality Control

We reformulate the four benchmarks' annotations without changing their semantics: clip or aggregated labels support question answering, annotated categories constrain video-based captions, and frame labels support temporal parsing. Instances from the same video share a partition, including videos appearing in both PanAf datasets. We filter generated responses for validity and consistency with the annotations. Appendix B details the conversion, splitting, and quality checks.

5 EthoBehavior: Compositional Ethogram Modeling

CEF provides a shared vocabulary for describing the meaning of behavior categories, but it does not by itself connect this structure to visual observations. We therefore develop **EthoBehavior**, which augments a pretrained video-language model with an **Ethogram Composition Adapter** (ECA). Given a video x, the visual encoder produces a sequence of visual tokens $\boldsymbol{H}_x$. ECA estimates the behavioral factors supported by $\boldsymbol{H}_x$ and represents them with two complementary forms of **Ethogram Factor Tokens** (EFTs): static factor tokens grounded in CEF and dynamic tokens learned from video data. The language model receives the original visual tokens together with both types of EFTs, allowing the same compositional representation to support behavior analysis. Figure 2 summarizes this information flow.

5.1 ETHOGRAM COMPOSITION ADAPTER

The eight factor vocabularies from Section 3 provide a common semantic space while each task retains the categories of its own ethogram. ECA associates each factor value with a learnable embedding $\boldsymbol{e}_{k,j}$ initialized from its textual description. The text initialization anchors the embedding in the meaning of the factor value, and training adapts it to visual cues. Appendix C.1 gives the parameterization.

To relate the video to these embeddings, ECA pools its visual tokens and incorporates the task context:

$$\boldsymbol{q}_{x,\tau} = \mathrm{Norm}\left(g(\mathrm{Pool}(\boldsymbol{H}_x)) + \boldsymbol{e}_\tau\right), \tag{5}$$

where $g(\cdot)$ is a lightweight projection network and $\boldsymbol{e}_\tau$ is the embedding of task type τ. The task embedding lets ECA emphasize evidence relevant to the requested analysis while reusing the same factor embeddings across tasks. The same video can therefore yield different factor probabilities for different tasks without changing the factor vocabulary. ECA then compares this video representation with each factor value:

$$s^{(k)}_{x,\tau,j} = \alpha\, \boldsymbol{q}^\top_{x,\tau} \boldsymbol{e}_{k,j} + b_{k,j}, \qquad p^{(k)}_{x,\tau,j} = \sigma\Big(s^{(k)}_{x,\tau,j}\Big), \tag{6}$$

where α is a learnable similarity scale and $b_{k,j}$ calibrates the score for each factor value. Independent sigmoid probabilities allow multiple values of one factor to be present when the video contains several animals or concurrent behaviors.

The targets and observation masks in Equation 4 indicate which factor values can be inferred from the annotations. Only observed entries contribute to the binary cross entropy loss:

$$\mathbb{O} = \{(i,k,j) \mid m^{(k)}_{i,j} = 1\}, \qquad \mathcal{L}_{\text{fac}} = \frac{1}{|\mathbb{O}|} \sum_{(i,k,j)\in\mathbb{O}} \mathrm{BCE}\Big(y^{(k)}_{i,j}, p^{(k)}_{x_i,\tau_i,j}\Big). \tag{7}$$

Here BCE denotes binary cross entropy, and $\mathbb{O}$ ranges over observed factor entries in the training data. Unspecified values contribute no loss and are never interpreted as negative examples. Because these targets describe the video as a whole, ECA does not assign factors to particular individuals or time intervals. The instruction supplies the relevant category vocabulary, so ECA does not require a unified classifier over all ethograms.

5.2 ETHOGRAM FACTOR TOKENS

Factor probabilities provide an explicit description of behavior content, but the language model operates on continuous token representations. EFTs connect these two spaces by converting the factor estimates produced by ECA into a compact sequence of video-conditioned embeddings.

Static factor tokens. Static EFTs are constructed from the named factor values in the fixed CEF vocabulary. Here, "static" refers to the fixed semantic basis rather than a video-independent representation: each token is selected and scaled according to its video-conditioned factor probability. For each behavior factor k, let $\mathcal{I}_{x,\tau,k}$ contain the $R = 2$ values with the highest predicted probabilities.

$$\boldsymbol{z}^s_{x,\tau,k,j} = g_\gamma\Big(p^{(k)}_{x,\tau,j}\Big)\left(\boldsymbol{W}_s \boldsymbol{e}_{k,j} + \boldsymbol{u}_k\right), \qquad g_\gamma(p) = p\,\mathbb{I}[p \geq \gamma], \tag{8}$$

where $\boldsymbol{W}_s$ projects the factor embedding to the language-model dimension, $\boldsymbol{u}_k$ preserves factor-type identity, and g_γ suppresses predictions with insufficient visual support. Each static EFT represents a factor value, its factor type, and its predicted confidence. Appendix C.2 clarifies how gradients pass through the selected tokens.

Dynamic residual tokens. The fixed factor vocabulary cannot name every visual detail. For each of the eight factors, we therefore learn a small dynamic codebook. A residual query contrasts the task-conditioned video representation with the factor estimates, selects one code, and passes it through a video-conditioned gate to produce one dynamic token. Unlike a static EFT, a dynamic residual token is not assigned to a named factor value. Appendix C.4 gives the query, code selection, straight-through training rule, and gate parameterization.

Table 1: Comparisons on PrInstruct-34K. Bold and underlined values indicate the best and second-best scores in each column, respectively.

Method	MQA				Behavior Captioning				Temporal Parsing		
	F1	Precision	Recall	Acc	BLEU-4	METEOR	ROUGE-L	CIDEr	F1@{0.3-0.7}	Edit	Frame-F1
Open-source Models											
GLM-5.3-Flash	27.95	28.12	28.94	23.73	19.37	21.31	38.03	**87.31**	9.04	29.20	<u>41.54</u>
Deepseek-V4.1-Flash	28.82	35.38	28.64	23.53	0.41	11.61	11.41	0	11.40	16.43	27.81
Video-Language Models											
InternVL3-8B	28.97	28.24	31.31	29.36	17.97	20.98	37.26	<u>77.07</u>	<u>13.91</u>	34.59	41.28
VideoLLaMA3-7B	24.15	24.97	25.11	17.43	13.76	18.35	32.78	63.41	8.28	26.14	35.17
LLaVAction-7B	<u>33.13</u>	31.84	<u>35.94</u>	24.57	17.37	19.90	36.61	68.84	13.67	<u>38.12</u>	**47.12**
Molmo2-4B	27.56	34.45	28.64	19.65	1.07	12.89	12.61	0	7.14	15.73	21.86
Molmo2-8B	28.27	33.50	32.70	11.96	0.97	12.52	12.50	0.01	13.15	15.00	22.16
Qwen3-VL-4B	23.21	23.75	24.26	16.40	14.66	18.85	33.96	70.52	8.93	25.34	33.35
EthoBehavior-4B	30.01	<u>37.22</u>	28.04	<u>31.95</u>	<u>26.84</u>	<u>26.98</u>	<u>47.19</u>	71.39	**14.42**	**39.99**	11.64
Qwen3-VL-8B	23.21	26.19	24.26	18.45	14.66	18.85	33.96	70.52	8.93	25.34	33.35
EthoBehavior-8B	**40.79**	**52.07**	**38.02**	**44.44**	**46.18**	**34.90**	**59.81**	72.68	10.32	36.45	23.06

Token injection and visual modulation. ECA also uses the factor predictions to modulate the original visual tokens through a low-rank residual transformation. Denote the resulting sequence by $\widetilde{\boldsymbol{H}}_{x,\tau}$; its construction is given in Appendix C.3. Let $\boldsymbol{Z}^{s}_{x,\tau}$ concatenate the selected explicit tokens in factor order, and let $\boldsymbol{Z}^{d}_{x,\tau}$ concatenate the dynamic tokens in the same order. They are inserted between the visual and instruction tokens:

$$\boldsymbol{H}^{\text{in}}_{x,\tau,\boldsymbol{r}} = \left[\widetilde{\boldsymbol{H}}_{x,\tau}; \boldsymbol{Z}^{s}_{x,\tau}; \boldsymbol{Z}^{d}_{x,\tau}; \boldsymbol{H}^{\text{inst}}_{\boldsymbol{r}}\right]. \tag{9}$$

The two EFT sequences contain 16 explicit and 8 dynamic tokens under the eight-factor specification. The unpooled visual sequence retains appearance and temporal order. In temporal parsing, the EFTs provide video-level context; the decoder must use the ordered visual tokens to identify interval boundaries and assign behaviors to those intervals.

5.3 Multi-task Learning

EthoBehavior uses one decoder for ethogram-conditioned question answering, behavior captioning, and temporal parsing, with each task expressed as conditional language generation. The generation objective $\mathcal{L}_{\text{gen}}$ averages token-level cross-entropy within each response and weights instances by ethogram–task group, limiting domination by long responses or frequent groups. The alignment objective $\mathcal{L}_{\text{align}}$ brings the task-conditioned video representation toward a frequency-balanced centroid of its confirmed positive factor embeddings; unspecified values are not treated as negatives.

We group the codebook diversity, and penalties into $\mathcal{L}_{\text{reg}}$. The overall training objective is:

$$\mathcal{L} = \mathcal{L}_{\text{gen}} + \lambda_{\text{fac}}\mathcal{L}_{\text{fac}} + \lambda_{\text{align}}\mathcal{L}_{\text{align}} + \mathcal{L}_{\text{reg}}. \tag{10}$$

We set $\lambda_{\text{fac}} = 0.25$ and $\lambda_{\text{align}} = 0.10$ for all experiments. Details are given in Appendix C.6.

6 Experiments

We conduct experiments to evaluate EthoBehavior from four complementary perspectives. First, we benchmark EthoBehavior with 4B and 8B backbones against general-purpose and animal behavior VLMs on PrInstruct-34K to assess its strong adaptability and behavior understanding capability. Second, we perform controlled ablations to examine the contributions of the ECA, EFT, partial factor supervision, and multi-task learning. Third, we evaluate action recognition on ChimpACT under the PriVi Mueller et al. (2026) protocol. Finally, we study generalization to unseen factor compositions and out-of-domain animal behaviors.

6.1 Implementation Details

We implement EthoBehavior on Qwen3-VL-4/8B Bai et al. (2025) and fine-tune the vision encoder, visual projector, and language model on 24,494 PrInstruct-34K training instances. Both variants use the same configuration for 10 epochs on two H200 GPUs, with a per-GPU batch size of 1 and eight gradient-accumulation steps. We uniformly sample 8-32 video frames at 4 fps during training and 16 frames for all models at evaluation. Additional detailed settings are given in Appendix C.7.

6.2 Evaluation Metrics

We evaluate each dataset in its native ethogram. For multi-label behavior question answering, we report accuracy and precision, recall, and F1. For behavior captioning, we report BLEU-4, METEOR, ROUGE-L, and CIDEr following the COCO caption protocol Chen et al. (2015). For temporal parsing, we report the mean segmental F1 over temporal IoU thresholds of 0.3, 0.5, and 0.7, normalized Edit score, and frame F1 Lea et al. (2017). For ChimpACT action recognition, we report mAP and weighted mAP Mueller et al. (2026). More metric rules are specified in Appendix D.

6.3 Results on PrInstruct-34K

Table 1 compares behavior question answering, captioning, and temporal parsing on PrInstruct-34K. The baselines span large general-purpose models GLM-5.3-Flash Z.ai (2026) and DeepSeek-V4.1-Flash Xu et al. (2026), video-language models InternVL3 Zhu et al. (2025), VideoLLaMA3 Zhang et al. (2025), Qwen3-VL Bai et al. (2025) and Molmo2 Clark et al. (2026), and the behavior-specific model LLaVAction Qi et al. (2026).

Overall, behavior-oriented VLMs can outperform much larger general-purpose multimodal LLMs on primate behavior understanding: the 7B LLaVAction exceeds GLM-5.3-Flash in MQA F1 (33.13 vs. 27.95) and both large models in temporal Edit. PrInstruct-34K makes this distinction consequential because models must identify categories from heterogeneous ethograms, describe the observed behavior, and recover its temporal structure rather than rely on generic video descriptions.

EthoBehavior improves substantially over its matched Qwen3-VL backbones in MQA, captioning, and segment-level temporal parsing, with a particularly large 8B gain in MQA F1 (40.79 vs. 23.21). Across model sizes, it achieves the best MQA F1, BLEU-4, METEOR, ROUGE-L, segmental F1, and Edit scores in Table 1, supporting the effectiveness of explicit ethogram composition for behavior understanding. Interestingly, EthoBehavior-4B exceeds the 8B variant on segmental F1 (14.42 vs. 10.32) and Edit despite its lower frame F1. This reversal suggests that stronger frame-level recognition in the larger model does not necessarily yield more coherent event boundaries and sequences, which temporal parsing explicitly evaluates. Representative captioning and temporal-parsing outputs, including remaining errors, are examined in Appendix G.

6.4 Component Ablations

Table 2 reports MQA F1 (QA), captioning BLEU-4 (Cap.), and mean segmental F1 (Temp.) for the 4B ablations. Removing ECA and EFTs causes the largest declines across all three tasks (6.14, 7.93, and 5.77 points), showing that the factor pathway contributes beyond the matched backbone, although this row cannot separate factor estimation from token injection. Visual modulation yields smaller but consistent gains even when EFTs are retained, suggesting that factor estimates are more useful when they also condition the visual sequence instead of reaching the decoder only as additional tokens.

Variant	QA	Cap.	Temp.
Full EthoBehavior-4B	**30.01**	26.84	14.42
Representation			
w/o ECA and EFTs	23.87	18.91	8.65
w/o static EFTs	27.42	23.96	12.31
w/o dynamic EFTs	28.16	24.88	12.76
w/o visual modulation	28.73	25.09	13.18
Supervision and training			
w/o partial factor loss	26.91	24.37	12.14
w/o alignment loss	29.24	25.58	14.31
Single-task training	28.26	24.97	**14.91**

Table 2: Component ablations on 4B backbone.

Removing the partial factor loss lowers all three metrics. Without direct supervision on observed factor values, EFTs and visual modulation depend on probabilities shaped indirectly by the remaining objectives, plausibly weakening their ethogram grounding. Training separate task-specific models reduces QA F1 and captioning BLEU-4 (30.01 vs. 28.26 and

26.84 vs. 24.97) but slightly improves temporal F1 (14.91 vs. 14.42). Joint training thus appears to help the two clip-level semantic tasks, whereas interval-level temporal parsing may have different optimization needs rather than benefiting uniformly from task sharing.

6.5 Generalization Analysis

Generalization to Unseen Factor Compositions. We construct three groups of unseen factor compositions pairing a predicate with participant role, posture, or direction values. For each group, training excludes examples containing the target pairing or any of its constituent factor values, while the corresponding test questions remain for evaluation. Because neither the combination nor its individual values receives task supervision, QA F1 on these questions probes transfer beyond memorized categories and factor labels. This is stricter than recombining factors seen during training, although it does not isolate value-level novelty from compositional novelty. EthoBehavior exceeds Qwen3-VL in all three groups (Appendix E, Table 6), with the largest gain on *move upward/downward* (36.36 vs. 20.85). These gains indicate that the factor-based representation retains useful structure even when the tested factor values and their combinations are absent from task training.

Generalization to Out-of-Domain Animal Behaviors. On Animal Kingdom Ng et al. (2022), zero-shot EthoBehavior-4B reaches 27.7 Jaccard Accuracy, close to the reported zero-shot InternVL3-8B score of 28.1 despite its smaller backbone (Appendix F, Table 7). This result suggests that its learned behavior representations retain some transferability beyond primates. However, the zero-shot score remains well below adapted references, including LLaVAction with InternVL3-8B at 58.7, so out-of-domain performance is not yet strong. One plausible reason is that Animal Kingdom's action definitions differ substantially from the primate-derived factor vocabulary, leaving some behaviors poorly represented by the learned composition. Because the reference methods use different training settings, these comparisons are indicative rather than strictly controlled.

6.6 Results on Action Recognition Benchmark

Table 3 evaluates subject-centric action recognition on ChimpACT under the protocol of Mueller et al. (2026), using mAP and weighted mAP rather than the generated outputs assessed on PrInstruct-34K. EthoBehavior-8B achieves the highest scores on both metrics, exceeding PriVi + DaLP by 3.17 mAP points and 6.82 weighted-mAP points. EthoBehavior-4B reaches 39.53 mAP, close to PriVi + DaLP at 40.00, but its weighted mAP is lower (56.07 vs. 59.29), so the advantage is not uniform across model sizes or metrics. This comparison is kept separate from Table 1 because ChimpACT uses subject-centric action labels and average precision, whereas PrInstruct-34K evaluates generated answers, captions, and temporal sequences. Matching the subject-box setting and category definitions is necessary for a meaningful comparison with the existing methods. These results show that EthoBehavior also performs competitively under a fair and subject-centric recognition protocol, beyond the multi-tasks instruction formats in Table 1.

	ChimpACT	
Method	mAP	$\mathrm{mAP_w}$
X3D (Feichtenhofer, 2020)	27.05	51.60
VideoMAEv2 (Wang et al., 2023)	–	–
UniFormerV2-B (Li et al., 2023)	–	–
InternVideo-L (Wang et al., 2022)	25.70	–
ChimpVLM (Brookes et al., 2024a)	–	–
VideoPrism-g (Zhao et al., 2024)	31.50	–
V-JEPA classifier (Bardes et al., 2024)	36.33	55.50
PriVi (Mueller et al., 2026)	39.25	58.16
PriVi + DaLP (Mueller et al., 2026)	40.00	59.29
EthoBehavior-4B	39.53	56.07
EthoBehavior-8B	**43.17**	**66.11**

Table 3: Action recognition on ChimpACT. $\mathrm{mAP_w}$ denotes weighted mAP.

7 Conclusion

We introduced CEF to connect heterogeneous primate ethograms through reusable factors while retaining their distinctions. PrInstruct-34K evaluates category identification, description, and temporal parsing, while EthoBehavior integrates partially supervised factor estimates as tokens into video-language modeling. Consistent gains over matched backbones and component ablations support the value of explicit composition. Factor Compositions tests show transfer beyond memorized labels, although they conflate unseen values and combinations. Zero-shot transfer beyond primates remains limited, motivating richer factor vocabularies for new ethograms.

AI USE STATEMENT

In this work, we used generative AI tools for generating synthetic datasets and polishing the writing of the manuscript. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

The factor representation and EthoBehavior are described in Sections 3 and 5, PrInstruct-34K construction and the experimental setup are summarized in Sections 4 and 6, and Appendices A–D provide the data processing steps, implementation settings, and evaluation protocols. The pretrained models and source benchmarks used in our experiments are public. We will publicly release the source code upon acceptance.

REFERENCES

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Max Bain, Arsha Nagrani, Daniel Schofield, Sophie Berdugo, Joana Bessa, Jake Owen, Kimberley J Hockings, Tetsuro Matsuzawa, Misato Hayashi, Dora Biro, et al. Automated audiovisual behavior recognition in wild primates. *Science advances*, 7(46):eabi4883, 2021.

Praneet C Bala, Benjamin R Eisenreich, Seng Bum Michael Yoo, Benjamin Y Hayden, Hyun Soo Park, and Jan Zimmermann. Automated markerless pose estimation in freely moving macaques with openmonkeystudio. *Nature communications*, 11(1):4560, 2020.

Adrien Bardes, Quentin Garrido, Jean Ponce, Xinlei Chen, Michael Rabbat, Yann LeCun, Mahmoud Assran, and Nicolas Ballas. Revisiting feature prediction for learning visual representations from video. *arXiv preprint arXiv:2404.08471*, 2024.

Gordon J Berman, Daniel M Choi, William Bialek, and Joshua W Shaevitz. Mapping the stereotyped behaviour of freely moving fruit flies. *Journal of The Royal Society Interface*, 11(99):20140672, 2014.

James P Bohnslav, Nivanthika K Wimalasena, Kelsey J Clausing, Yu Y Dai, David A Yarmolinsky, Tomás Cruz, Adam D Kashlan, M Eugenia Chiappe, Lauren L Orefice, Clifford J Woolf, et al. Deepethogram, a machine learning pipeline for supervised behavior classification from raw pixels. *elife*, 10:e63377, 2021.

Otto Brookes, Majid Mirmehdi, Hjalmar Kuhl, and Tilo Burghardt. Chimpvlm: Ethogram-enhanced chimpanzee behaviour recognition. *arXiv preprint arXiv:2404.08937*, 2024a.

Otto Brookes, Majid Mirmehdi, Colleen Stephens, Samuel Angedakin, Katherine Corogenes, Dervla Dowd, Paula Dieguez, Thurston C Hicks, Sorrel Jones, Kevin Lee, et al. Panaf20k: A large video dataset for wild ape detection and behaviour recognition. *International Journal of Computer Vision*, 132(8):3086–3102, 2024b.

Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015.

Christopher Clark, Jieyu Zhang, Zixian Ma, Jae Sung Park, Rohun Tripathi, Sangho Lee, Mohammadreza Salehi, Jason Ren, Chris Dongjoo Kim, Yinuo Yang, et al. Molmo2: Open weights and data for vision-language models with video understanding and grounding. In *Conference on Computer Vision and Pattern Recognition*, pp. 28652–28668, 2026.

Isla Duporge, Maksim Kholiavchenko, Roi Harel, Scott Wolf, Daniel I Rubenstein, Margaret C Crofoot, Tanya Berger-Wolf, Stephen J Lee, Julie Barreau, Jenna Kline, et al. Baboonland dataset: Tracking primates in the wild and automating behaviour recognition from drone videos: I. duporge et al. *International Journal of Computer Vision*, 133(9):6578–6589, 2025.

Christoph Feichtenhofer. X3d: Expanding architectures for efficient video recognition. In *Conference on Computer Vision and Pattern Recognition*, pp. 200–210, 2020.

Eliaf Y Garcia-Loya, Irvin Hussein López-Nava, and Verónica Reyes-Meza. What are we measuring? differentiating animal actions, activities, activity level, and behavior in computational ethology. *Applied Animal Behaviour Science*, pp. 107014, 2026.

Georgios V Gkoutos, Paul N Schofield, and Robert Hoehndorf. The neurobehavior ontology: an ontology for annotation and integration of behavior and behavioral phenotypes. *International review of neurobiology*, 103:69–87, 2012.

Georgios V Gkoutos, Robert Hoehndorf, Loukia Tsaprouni, and Paul N Schofield. Best behaviour? ontologies and the formal description of animal behaviour. *Mammalian Genome*, 26(9):540–547, 2015.

Alexander I Hsu and Eric A Yttri. B-soid, an open-source unsupervised algorithm for identification and fast prediction of behaviors. *Nature communications*, 12(1):5188, 2021.

Yinuo Jing, Jinyan Wu, Zixi Yang, Kongming Liang, Xiatian Zhu, and Zhanyu Ma. Ethoclip: Ontology-enhanced video-language pretraining for animal behavior understanding. In *Conference on Computer Vision and Pattern Recognition*, pp. 31196–31206, 2026.

Mayank Kabra, Alice A Robie, Marta Rivera-Alba, Steven Branson, and Kristin Branson. Jaaba: interactive machine learning for automatic annotation of animal behavior. *Nature methods*, 10(1): 64–67, 2013.

Colin Lea, Michael D. Flynn, René Vidal, Austin Reiter, and Gregory D. Hager. Temporal convolutional networks for action segmentation and detection. In *Conference on Computer Vision and Pattern Recognition*, pp. 1003–1012, 2017.

Daniel A Levitis, William Z Lidicker Jr, and Glenn Freund. Behavioural biologists do not agree on what constitutes behaviour. *Animal behaviour*, 78(1):103–110, 2009.

Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Limin Wang, and Yu Qiao. Uniformerv2: Unlocking the potential of image vits for video understanding. pp. 1632–1643, 2023.

Hongli Liu, Yu Wang, and Shengjie Zhao. Star: Semantic-temporal adaptive representation learning for few-shot action recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, 2026.

Kevin Luxem, Petra Mocellin, Falko Fuhrmann, Johannes Kürsch, Stephanie R Miller, Jorge J Palop, Stefan Remy, and Pavol Bauer. Identifying behavioral structure from deep variational embeddings of animal motion. *Communications Biology*, 5(1):1267, 2022.

Xiaoxuan Ma, Stephan Kaufhold, Jiajun Su, Wentao Zhu, Jack Terwilliger, Andres Meza, Yixin Zhu, Federico Rossano, and Yizhou Wang. Chimpact: A longitudinal dataset for understanding chimpanzee behaviors. volume 36, pp. 27501–27531, 2023.

Joanna Materzynska, Tete Xiao, Roei Herzig, Huijuan Xu, Xiaolong Wang, and Trevor Darrell. Something-else: Compositional action recognition with spatial-temporal interaction networks. In *Conference on Computer Vision and Pattern Recognition*, pp. 1046–1056, 2020.

Alexander Mathis, Pranav Mamidanna, Kevin M Cury, Taiga Abe, Venkatesh N Murthy, Mackenzie Weygandt Mathis, and Matthias Bethge. Deeplabcut: markerless pose estimation of user-defined body parts with deep learning. *Nature neuroscience*, 21(9):1281–1289, 2018.

Felix B Mueller, Jan F Meier, Timo Lueddecke, Richard Vogg, Roger L Freixanet, Valentin Hassler, Tiffany Bosshard, Elif Karakoc, William J O'Hearn, Sofia M Pereira, et al. Privi: Towards a general-purpose video model for primate behavior in the wild. In *Conference on Computer Vision and Pattern Recognition*, pp. 38804–38815, 2026.

Xun Long Ng, Kian Eng Ong, Qichen Zheng, Yun Ni, Si Yong Yeo, and Jun Liu. Animal kingdom: A large and diverse dataset for animal behavior understanding. In *Conference on Computer Vision and Pattern Recognition*, pp. 19001–19012, 2022.

Talmo D Pereira, Nathaniel Tabris, Arie Matsliah, David M Turner, Junyu Li, Shruthi Ravindranath, Eleni S Papadoyannis, Edna Normand, David S Deutsch, Z Yan Wang, et al. Sleap: A deep learning system for multi-animal pose tracking. *Nature methods*, 19(4):486–495, 2022.

Haozhe Qi, Shaokai Ye, Alexander Mathis, and Mackenzie Mathis. Llavaction: evaluating and training multi-modal large language models for action understanding. In *International Conference on Learning Representations*, volume 2026, pp. 21859–21883, 2026.

Arka Sadhu, Tanmay Gupta, Mark Yatskar, Ram Nevatia, and Aniruddha Kembhavi. Visual semantic role labeling for video understanding. In *Conference on Computer Vision and Pattern Recognition*, pp. 5585–5596, 2021.

Paulina M Schenk, Robert West, Oscar Castro, Emily Hayes, Janna Hastings, Marie Johnston, Marta M Marques, Elizabeth Corker, Alison J Wright, Gabriella Stuart, et al. An ontological framework for organising and describing behaviours: The human behaviour ontology. *Wellcome Open Research*, 9:237, 2025.

Qwen Team. Qwen3.5: Towards native multimodal agents, 2026. URL https://qwen.ai/blog?id=qwen3.5.

Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. volume 30, 2017.

Richard Vogg, Timo Lueddecke, Jonathan Henrich, Sharmita Dey, Matthias Nuske, Valentin Hassler, Derek Murphy, Julia Fischer, Julia Ostner, Oliver Schuelke, et al. Computer vision for primate behavior analysis in the wild. *Nature Methods*, 22(6):1154–1166, 2025.

Limin Wang et al. VideoMAE V2: Scaling video masked autoencoders with dual masking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. URL https://arxiv.org/abs/2303.16727.

Yi Wang, Kunchang Li, Yizhuo Li, Yinan He, Bingkun Huang, Zhiyu Zhao, Hongjie Zhang, Jilan Xu, Yi Liu, Zun Wang, et al. Internvideo: General video foundation models via generative and discriminative learning. *arXiv preprint arXiv:2212.03191*, 2022.

Caleb Weinreb, Jonah E Pearl, Sherry Lin, Mohammed Abdal Monium Osman, Libby Zhang, Sidharth Annapragada, Eli Conlin, Red Hoffmann, Sofia Makowska, Winthrop F Gillis, et al. Keypoint-moseq: parsing behavior by linking point tracking to pose dynamics. *Nature methods*, 21(7):1329–1339, 2024.

Alexander B Wiltschko, Matthew J Johnson, Giuliano Iurilli, Ralph E Peterson, Jesse M Katon, Stan L Pashkovski, Victoria E Abraira, Ryan P Adams, and Sandeep Robert Datta. Mapping sub-second structure in mouse behavior. *Neuron*, 88(6):1121–1135, 2015.

Anyi Xu, B Li, Bangcai Lin, Bing Xue, BingCheng Xian, Bingzheng Xu, Bochao Wu, Bowei Zhang, Boyi Deng, CC Yu, et al. Deepseek-v4.1-flash: Pushing the limits of kv cache compression. *arXiv preprint arXiv:2609.19969*, 2026.

Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

Yuan Yao, Praneet Bala, Abhiraj Mohan, Eliza Bliss-Moreau, Kristine Coleman, Sienna M Freeman, Christopher J Machado, Jessica Raper, Jan Zimmermann, Benjamin Y Hayden, et al. Openmonkeychallenge: dataset and benchmark challenges for pose estimation of non-human primates. *International journal of computer vision*, 131(1):243–258, 2023.

Mark Yatskar, Luke Zettlemoyer, and Ali Farhadi. Situation recognition: Visual semantic role labeling for image understanding. In *Conference on Computer Vision and Pattern Recognition*, pp. 5534–5542, 2016.

Z.ai. Glm-5.3-flash: Frontier intelligence, flash cost, 2026. URL https://z.ai/blog/glm-5.3-flash.

Boqiang Zhang, Kehan Li, Zesen Cheng, Zhiqiang Hu, Yuqian Yuan, Guanzheng Chen, Sicong Leng, Yuming Jiang, Hang Zhang, Xin Li, et al. Videollama 3: Frontier multimodal foundation models for image and video understanding. *arXiv preprint arXiv:2501.13106*, 2025.

Long Zhao, Nitesh B Gundavarapu, Liangzhe Yuan, Hao Zhou, Shen Yan, Jennifer J Sun, Luke Friedman, Rui Qian, Tobias Weyand, Yue Zhao, et al. Videoprism: A foundational visual encoder for video understanding. 2024.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025.

A Partial Factor Supervision Details

For an instance with a category vector covering all of $\mathbb{C}_{d_i}$, let $\mathbb{N}_i^{\text{cat}}$ contain its negative categories and set $\chi_i = 1$; otherwise set $\chi_i = 0$ and $\mathbb{N}_i^{\text{cat}} = \emptyset$. For factor value $u \in \mathbb{U}$, the categories in ethogram d that could express it are

$$\text{Cand}_d(u) = \{(d, \ell) : \ell \in \mathbb{C}_d,\ u \in \text{Poss}((d, \ell))\}. \tag{11}$$

The confirmed negative set used in Equation 4 is

$$\mathbb{N}_i = \left\{u \in \mathbb{U} \setminus \mathbb{F}_i : \chi_i = 1,\ \text{Cand}_{d_i}(u) \neq \emptyset,\ \text{Cand}_{d_i}(u) \subseteq \mathbb{N}_i^{\text{cat}}\right\}. \tag{12}$$

Thus a value is not marked negative merely because it is omitted from a positive category. The rule assumes that a complete category vector covers the relevant behavior in the video; when individuals or intervals are unannotated, its inferred negatives may be unreliable.

B PrInstruct-34K Construction Details

Each video contributes instances only to tasks supported by its annotation type. For multiple-choice question answering, we use the clip-level labels of PanAf20K and aggregate frame- or instance-level annotations over time and individuals for the other benchmarks. For behavior captioning, Qwen3.5-397B Team (2026) generates a concise description conditioned on the video, its annotated categories, and their definitions. The categories constrain behavioral content while the video provides visual context. For temporal parsing, frame-level annotations from PanAf500, BaboonLand, and ChimpACT are converted into active behavior sets. Consecutive frames with identical sets are merged, and their boundaries are converted to seconds.

Table 4: Composition of PrInstruct-34K. Counts denote instruction instances; caption counts report generations retained after filtering.

Source	Categories	MQA	Caption	Temporal	Total
PanAf20K (Brookes et al., 2024b)	18	15,633	14,225	–	29,858
PanAf500 (Brookes et al., 2024b)	9	500	371	500	1,371
BaboonLand (Duporge et al., 2025)	13	1,066	783	1,066	2,915
ChimpACT (Ma et al., 2023)	23	163	163	156	482
PrInstruct-34K	58	17,362	15,542	1,722	34,626

All task instances derived from the same video are assigned to the same train, validation, or test partition. Since PanAf20K and PanAf500 contain overlapping videos, we repartition the data using SHA-256 hashes to make the resulting splits mutually disjoint.

We apply automatic checks for response validity and consistency with the benchmark annotations. A generation is rejected if it cannot be parsed as a structured response, contains an empty caption, introduces a category outside the annotated behavior set, or mentions an unannotated category detected through an alias-aware lexical scan. Among 17,362 generated caption records, the filter retained 15,542 and rejected 1,820, an overall rejection rate of 10.5%. The rejection rates were 9.0% for PanAf20K, 25.8% for PanAf500, and 26.5% for BaboonLand. All ChimpACT records produced a valid response after constrained regeneration.

C EthoBehavior Implementation Details

C.1 Factor Embeddings

For value $v_{k,j}$, we tokenize its factor key (replacing `::` with a space), followed by a colon and its description, using the tokenizer of the corresponding Qwen3-VL backbone. Let $u_{k,j,1}, \ldots, u_{k,j,L_{k,j}}$ be these tokens and let Emb_{Qwen} denote the backbone's input embedding table.

The fixed text initialization is obtained by mean pooling and a non-trainable random projection into the $d_o = 256$ dimensional factor space:

$$\boldsymbol{t}_{k,j} = \frac{1}{L_{k,j}} \sum_{\ell=1}^{L_{k,j}} \mathrm{Emb}_{\mathrm{Qwen}}(u_{k,j,\ell}), \qquad \boldsymbol{e}^{\mathrm{text}}_{k,j} = \mathrm{Norm}(\boldsymbol{P}^\top \boldsymbol{t}_{k,j}). \tag{13}$$

Here $\boldsymbol{P} \in \mathbb{R}^{D_{\mathrm{Qwen}} \times d_o}$ has entries sampled from $\mathcal{N}(0, 1/d_o)$ with seed 17, and the projected text vectors are computed once before training. Each factor-value embedding then adds a trainable factor-type embedding and a trainable value residual:

$$\boldsymbol{e}_{k,j} = \mathrm{Norm}\big(\boldsymbol{e}^{\mathrm{text}}_{k,j} + \boldsymbol{e}^{\mathrm{type}}_{k} + \Delta \boldsymbol{e}_{k,j}\big)\,. \tag{14}$$

The residual $\Delta \boldsymbol{e}_{k,j}$ is initialized to zero; no separate pretrained text encoder is used.

C.2 Static Factor Token Selection

Top-R selection uses hard indices, so gradients from the language loss update selected probabilities and embeddings but not the choice of indices. Thresholding zeroes unsupported embeddings without shortening the token sequence.

C.3 Visual-Token Modulation

The modulation branch uses all task-conditioned factor probabilities, without Top-R selection or thresholding, to form a global factor context:

$$\boldsymbol{c}_{x,\tau} = \frac{\sum_{k\in\mathbb{K}} \sum_{j=1}^{n_k} p^{(k)}_{x,\tau,j} \boldsymbol{e}_{k,j}}{\sum_{k\in\mathbb{K}} \sum_{j=1}^{n_k} p^{(k)}_{x,\tau,j} + \epsilon}. \tag{15}$$

For a visual token $\boldsymbol{h}_{x,t}$, a gate from this context modulates a rank-r transformation:

$$\begin{aligned} \boldsymbol{a}_{x,\tau} &= 2\sigma(\boldsymbol{W}_g \boldsymbol{c}_{x,\tau} + \boldsymbol{b}_g), \\ \Delta \boldsymbol{h}_{x,\tau,t} &= \boldsymbol{W}_{\mathrm{up}}(\mathrm{GELU}(\boldsymbol{W}_{\mathrm{down}}\,\mathrm{LN}(\boldsymbol{h}_{x,t})) \odot \boldsymbol{a}_{x,\tau})\,, \\ \widetilde{\boldsymbol{h}}_{x,\tau,t} &= \boldsymbol{h}_{x,t} + \Delta \boldsymbol{h}_{x,\tau,t}, \end{aligned} \tag{16}$$

where $r = 64$, $\boldsymbol{W}_{\mathrm{down}}$ maps to r dimensions, and $\boldsymbol{W}_{\mathrm{up}}$ maps back to the language-model dimension. The branch applies only to video-token positions, leaving instruction and other text tokens unchanged. The video bypass scale is one and modulation dropout is zero in both model variants. The gate parameters $\boldsymbol{W}_g, \boldsymbol{b}_g$ and output projection $\boldsymbol{W}_{\mathrm{up}}$ are zero-initialized, so the modulation initially acts as the identity.

C.4 Dynamic Codebook and Token Construction

For each factor k, the residual branch first summarizes the predicted factor values, including those outside the Top-R set:

$$\bar{\boldsymbol{e}}_{x,\tau,k} = \frac{\sum_{j=1}^{n_k} p^{(k)}_{x,\tau,j} \boldsymbol{e}_{k,j}}{\sum_{j=1}^{n_k} p^{(k)}_{x,\tau,j} + \epsilon}. \tag{17}$$

A learned factor-specific vector $\boldsymbol{u}^r_k$ distinguishes the residual query:

$$\boldsymbol{u}^r_{x,\tau,k} = \mathrm{Norm}(\boldsymbol{q}_{x,\tau} + \boldsymbol{u}^r_k), \qquad \boldsymbol{r}_{x,\tau,k} = \mathrm{Norm}\big(f_r\left[\boldsymbol{u}^r_{x,\tau,k}; \bar{\boldsymbol{e}}_{x,\tau,k}; \boldsymbol{u}^r_{x,\tau,k} - \bar{\boldsymbol{e}}_{x,\tau,k}\right]\big)\,. \tag{18}$$

The codebook for factor k is $\mathbb{R}_k = \{\boldsymbol{b}_{k,1}, \ldots, \boldsymbol{b}_{k,C}\}$ with $C = 4$ and $\widehat{\boldsymbol{b}}_{k,h} = \mathrm{Norm}(\boldsymbol{b}_{k,h})$. Code scores use temperature $T_d = 0.1$, and their maximum determines the hard selection:

$$\zeta_{x,\tau,k,h} = \boldsymbol{r}^\top_{x,\tau,k} \widehat{\boldsymbol{b}}_{k,h} / T_d, \qquad h^*_{x,\tau,k} = \operatorname*{arg\,max}_{1\le h\le C} \zeta_{x,\tau,k,h}. \tag{19}$$

Let $\pi_{x,\tau,k,h}$ be the softmax of these scores. Inspired by discrete-codebook learning (Van Den Oord et al., 2017), we use hard-forward, soft-backward selection: the forward pass selects $\widehat{\boldsymbol{b}}_{k,h^*_{x,\tau,k}}$,

while gradients pass through the softmax probabilities $\pi_{x,\tau,k,h}$. We denote this hard-forward, soft-backward code vector by $\widetilde{\boldsymbol{b}}_{x,\tau,k}$. A video-conditioned gate scales the projected code to form the dynamic token:

$$\rho_{x,\tau,k} = \sigma(f_g[\boldsymbol{r}_{x,\tau,k}; \bar{\boldsymbol{e}}_{x,\tau,k}]), \qquad \boldsymbol{z}^d_{x,\tau,k} = \rho_{x,\tau,k}\left(\boldsymbol{W}_d\widetilde{\boldsymbol{b}}_{x,\tau,k} + \boldsymbol{u}_k\right). \tag{20}$$

The factor-type embedding $\boldsymbol{u}_k$ is shared by the static and dynamic EFTs, placing both token types in the same factor-specific subspace.

C.5 Dynamic Regularization

Here sg denotes stop-gradient. With expectations denoting uniform averages over the displayed instances, factors, codes, and values, respectively, the four penalties are

$$\begin{aligned}\mathcal{L}_{\text{com}} &= \mathbb{E}_{i,k}\left[1 - \widetilde{\boldsymbol{b}}^\top_{x_i,\tau_i,k}\,\text{sg}(\boldsymbol{r}_{x_i,\tau_i,k})\right] + \beta\mathbb{E}_{i,k}\left[1 - \boldsymbol{r}^\top_{x_i,\tau_i,k}\,\text{sg}(\widetilde{\boldsymbol{b}}_{x_i,\tau_i,k})\right],\\ \mathcal{L}_{\text{orth}} &= \mathbb{E}_{k,h,j}\left[(\widehat{\boldsymbol{b}}^\top_{k,h}\boldsymbol{e}_{k,j})^2\right],\end{aligned} \tag{21}$$

where $\beta = 0.25$, and

$$\begin{aligned}\mathcal{L}_{\text{div}} &= \mathbb{E}_{k,h\neq h'}\left[(\widehat{\boldsymbol{b}}^\top_{k,h}\widehat{\boldsymbol{b}}_{k,h'})^2\right],\\ \mathcal{L}_{\text{spa}} &= \mathbb{E}_{i,k}[\rho_{x_i,\tau_i,k}].\end{aligned} \tag{22}$$

The diversity term separates code vectors geometrically; it does not enforce equal code use. The grouped regularizer in the main objective is

$$\mathcal{L}_{\text{reg}} = \lambda_{\text{com}}\mathcal{L}_{\text{com}} + \lambda_{\text{orth}}\mathcal{L}_{\text{orth}} + \lambda_{\text{div}}\mathcal{L}_{\text{div}} + \lambda_{\text{spa}}\mathcal{L}_{\text{spa}}, \tag{23}$$

with $\lambda_{\text{com}} = 0.05$, $\lambda_{\text{orth}} = \lambda_{\text{div}} = 0.01$, and $\lambda_{\text{spa}} = 0.005$.

C.6 Generation and Alignment Objectives

For each instance i in a training batch $\mathbb{B}$, $\mathcal{L}_{\text{gen}}$ averages teacher-forced token cross-entropy over its response length and applies an ethogram–task group weight w_{d_i,τ_i} before averaging across instances. Here d_i and τ_i identify the ethogram and task. Each source benchmark supplies one ethogram in our four-benchmark setup, so the implemented group counts can be indexed by ethogram–task pairs. Let $n_{d,\tau}$ be the number of training instances in a nonempty pair, $N = \sum_{d,\tau} n_{d,\tau} = 24{,}494$, and $\mathbb{G}_+ = \{(d,\tau) : n_{d,\tau} > 0\}$. With inverse-frequency exponent one, the 11 nonempty groups receive weights

$$w_{d,\tau} = \frac{N}{|\mathbb{G}_+|\, n_{d,\tau}}, \qquad (d,\tau) \in \mathbb{G}_+. \tag{24}$$

Let $\mathbb{P}_i = \{(k,j) \mid k \in \mathbb{K},\ 1 \le j \le n_k,\ m^{(k)}_{i,j} = 1,\ y^{(k)}_{i,j} = 1\}$ contain known positive factors. Let $c_{k,j}$ count training instances in which value (k,j) is a confirmed positive, and let $K_+ = |\{(k,j) : c_{k,j} > 0\}| = 39$. The frequency-balancing weight is

$$\omega_{k,j} = \begin{cases} N/(K_+ c_{k,j}), & c_{k,j} > 0,\\ 0, & c_{k,j} = 0.\end{cases} \tag{25}$$

Their frequency-balanced embedding centroid is

$$\bar{\boldsymbol{e}}_i = \text{Norm}\left(\sum_{(k,j)\in\mathbb{P}_i} \omega_{k,j}\boldsymbol{e}_{k,j}\right). \tag{26}$$

For $\mathbb{B}^+ = \{i \in \mathbb{B} \mid |\mathbb{P}_i| > 0\}$, the alignment objective is

$$\mathcal{L}_{\text{align}} = \frac{1}{|\mathbb{B}^+|}\sum_{i\in\mathbb{B}^+}\left(1 - \cos(\boldsymbol{q}_{x_i,\tau_i}, \bar{\boldsymbol{e}}_i)\right). \tag{27}$$

Only confirmed positives contribute; unspecified values are not used as contrastive negatives. The alignment term is omitted when $\mathbb{B}^+ = \emptyset$.

C.7 Training and Evaluation Configuration

Table 5 lists settings shared by the 4B and 8B variants. Visibility remains a separate observation variable and is not counted as a ninth action factor.

Table 5: Training and evaluation settings for EthoBehavior.

Training	
Epochs	10
Batch size / GPU	1
Gradient accumulation	8
Learning rate	2×10^{-6}
LR schedule	cosine
Warm-up ratio	0.03
Train frames	8-32 at 4 fps
Adapter	
ECA hidden dimension	256
Static selection	top-2 / factor
Static threshold	0.15
Visual modulation rank	64
Evaluation	
Video frames	16 segment centers

Training uses DeepSpeed ZeRO-3 and gradient checkpointing, with no weight decay, a maximum gradient norm of 1, and a maximum input length of 8,192 tokens. The language modeling loss has unit weight, with $\lambda_{\text{fac}} = 0.25$ and $\lambda_{\text{align}} = 0.10$.

For evaluation, we uniformly divide each video into 16 temporal segments and sample the center frame of each segment. API-based models receive these frames and their timestamps as an ordered image sequence. All models receive the same task instructions and output formats, and predictions are processed by the same automatic parser without manual correction. We use each dataset's test partition, except that BaboonLand's validation partition serves as its evaluation set and is never used for optimization. Temporal parsing is evaluated only on datasets with temporal behavior annotations.

D Evaluation Protocols

We compute every metric separately for each dataset in its native category space. Question-answering, temporal-parsing, and action-recognition scores are reported as percentages. Captioning scores returned by the MS COCO evaluation toolkit Chen et al. (2015) are multiplied by 100 to match the result tables.

Behavior question answering. Exact-set accuracy counts an answer as correct only when its predicted category set equals the annotated set. Macro precision, recall, and F1 average the respective per-category scores, giving each category equal weight despite class imbalance.

Behavior captioning. BLEU-4, METEOR, ROUGE-L, and CIDEr are computed using the MS COCO evaluator with Stanford PTB tokenization Chen et al. (2015).

Temporal parsing. Predicted and annotated behavior labels at each frame are treated as sets because several behaviors may overlap. Frame F1 is computed for each category over all evaluated frames and then macro-averaged. For normalized Edit score, consecutive frames with identical active-label sets are collapsed, and the score equals one minus the Levenshtein distance between the predicted and reference state sequences divided by the length of the longer sequence. For segmental F1, maximal contiguous regions are extracted separately for each category. Predicted and reference segments of the same category are matched one-to-one when their temporal intersection over union reaches 0.3, 0.5, or 0.7. We macro-average F1 across categories at each threshold and then report the mean over the three thresholds, adapting the segmental evaluation of Lea et al. (2017) to multi-label behavior sequences.

ChimpACT action recognition. Following the PriVi protocol Mueller et al. (2026), average precision uses ground-truth subject boxes, spatial matching at IoU 0.5, and the VOC precision envelope. The reported mAP averages category AP values uniformly, whereas support-weighted mAP weights each category by its number of ground-truth instances.

E Unseen Factor Composition Results

Table 6 reports QA F1 for three groups of unseen factor compositions. Training excludes each target composition and every example containing its constituent factor values; evaluation uses the corresponding questions.

Unseen factor composition	EthoBehavior	Qwen3-VL
contact actor/recipient	20.90	16.17
maintain sitting/standing	20.33	16.24
move upward/downward	36.36	20.85

Table 6: QA F1 on unseen factor compositions. Higher is better.

EthoBehavior improves over Qwen3-VL in every unseen factor composition group, by 4.73 F1 points for *contact actor/recipient*, 4.09 for *maintain sitting/standing*, and 15.51 for *move upward/downward*. The largest gain occurs for movement direction, suggesting that the shared factor representation remains useful when the target pairing and its values are absent from training. However, F1 remains near 20 for the role and posture groups, so transfer is uneven rather than solved by the factor representation. Because training also excludes the constituent factor values, these results test unseen factor compositions under value-level novelty and do not isolate recombination of previously seen values.

F Out-of-Domain Animal Behavior Results

Table 7 reports Jaccard accuracy on the multi-label Animal Kingdom action-recognition task Ng et al. (2022). The reference scores follow Qi et al. (2026).

Method	Jaccard Acc.
VideoMAE	53.1
LLaVA-Video-7B (zero-shot)	30.5
LLaVAction (LLaVA-Video-7B)	61.0
InternVL3-8B (zero-shot)	28.1
LLaVAction (InternVL3-8B)	58.7
EthoBehavior-4B (zero-shot)	27.7

Table 7: Action recognition on Animal Kingdom.

EthoBehavior-4B attains 27.7 Jaccard accuracy, 0.4 points below zero-shot InternVL3-8B and 2.8 points below zero-shot LLaVA-Video-7B. This places a smaller primate-trained model close to the two larger zero-shot VLMs, but it does not establish an out-of-domain advantage. VideoMAE and the two LLaVAction variants score substantially higher (53.1–61.0), leaving a clear transfer gap. Because the compared systems differ in backbones and training protocols, this table cannot isolate whether that gap is due to a mismatch between the primate-derived factors and Animal Kingdom classes. Adapting the factor definitions to the new ethogram and testing under matched supervision would help assess this explanation.

G Qualitative Results

The following examples complement the aggregate results in Table 1 by showing model outputs for behavior captioning and temporal parsing.

In the chimpanzee example, EthoBehavior describes the seated animal and its object interaction more conservatively than baselines that introduce grooming or foraging, but it omits the second chimpanzee entering the scene. In the baboon example, it retains group posture and movement

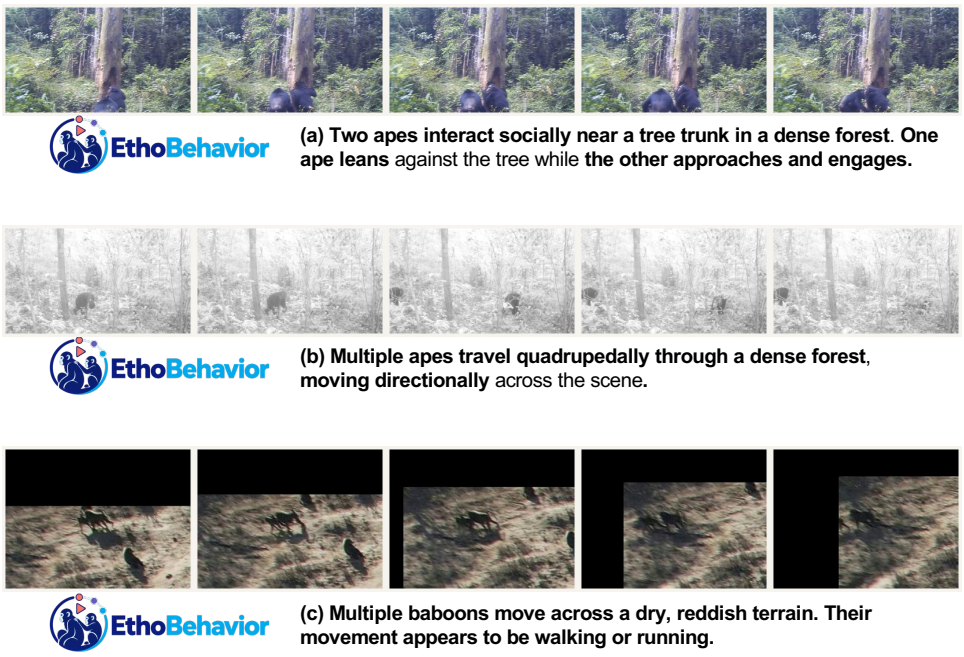


Figure 3: Additional EthoBehavior captions for primate interactions and group locomotion across three clips. These outputs illustrate the behavioral content of the generated descriptions.

across rocky terrain whereas DeepSeek-V4.1-Flash describes a solitary animal, although its additional foraging detail is not established by the ground-truth description.

The captions express participant number, interaction, locomotion mode, and movement direction across chimpanzee and baboon scenes. These examples illustrate the range of compositional descriptions rather than establish a separate performance claim.

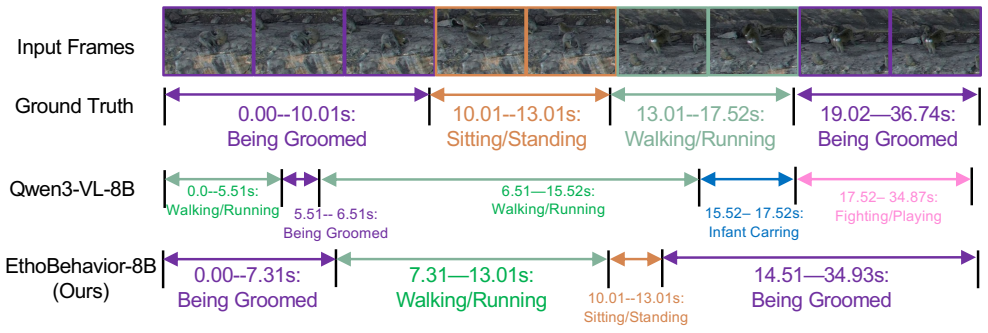
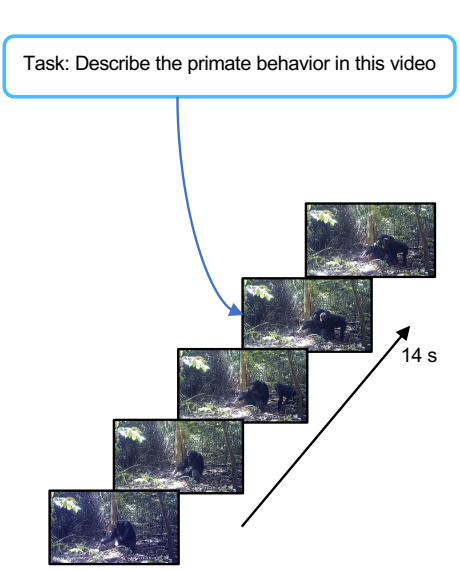


Figure 5: Temporal-parsing example comparing the ground truth with Qwen3-VL-8B and EthoBehavior-8B. Intervals may overlap because the task allows multiple active behavior labels.

EthoBehavior recovers grooming near the beginning and end of the clip and avoids the baseline's unsupported infant-carrying and fighting predictions. It still predicts walking too early and misaligns the middle transitions, showing why improved behavior identification does not automatically yield precise temporal boundaries.



GT Annotation

"A chimpanzee sits on the forest floor interacting with an object while another chimpanzee walks into the frame and stands nearby."

DeepSeek-V4.1-Flash

There are **two distinct subjects** visible in the video: a large, dark-furred ape (**likely a chimpanzee**) and a smaller, dark-furred ape (**likely an infant or juvenile**). A third, **smaller dark shape appears on the far right edge of the frame starting at timestamp 7.917 seconds, but it is partially obscured and only visible in a few frames, making it difficult to confirm if it is a third individual or just a shadow/part of the environment.**

Molmo2-8B

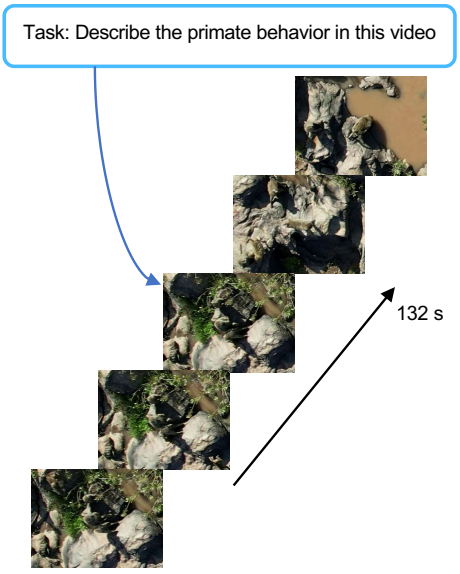
In this video, I observe **two chimpanzees** displaying distinct behaviors in their natural forest habitat.
The first chimpanzee is the primary subject, **sitting on the forest floor** with its back to the camera. This individual is **engaged in foraging behavior**, using its right hand to pick up and **manipulate objects** from the ground. The chimpanzee appears focused on its task, **repeatedly reaching down to grasp items and then bringing them to its mouth. This suggests the animal is either eating or examining something it has found.**
The second chimpanzee **enters the scene** from the right side of the frame. This individual **approaches the seated chimpanzee and stands nearby. The standing chimpanzee then leans forward and places its right hand on the back of the seated chimpanzee.**

Qwen3-VL-8B

Two chimpanzees interact in a forest; **one grooms** the other while **they are both bent over on the ground.**

EthoBehavior-8B (Ours)

A chimpanzee sits on a log **in a forest, interacting with something on the ground while remaining seated.**



GT Annotation

"Baboons are seen moving across rocky terrain, some sitting or standing on rocks, while others appear to groom themselves or drink water near a stream"

DeepSeek-V4.1-Flash

There is a single, solitary baboon visible throughout the sequence. It appears to be an adult or large sub-adult, given its size relative to the rocks and vegetation. **There are no other baboons or animals visible in the frames.**

Molmo2-8B

The video shows **a group of baboons** exhibiting **typical social and foraging behaviors** in their natural habitat. **The group consists of multiple individuals, with at least one clearly visible baboon in the foreground and others partially visible in the background.**
The baboons display a range of postures and movements characteristic of their species. **Some individuals are standing upright, while others are sitting** or lying down. **The baboons are actively engaged in foraging behavior, with several individuals using their hands to pick up and manipulate objects. One baboon is particularly noticeable as it holds an object in its hand, demonstrating the dexterous use of their hands for food handling. Social interaction is evident among the group members. They appear to be moving together in a coordinated manner, suggesting a cohesive social structure.**

Qwen3-VL-8B

Baboons move across rocky terrain near water, with some individuals sitting or standing while others **engage in physical interaction.**

EthoBehavior-8B (Ours)

Baboons are seen sitting, standing, and walking on rocks near a muddy water body with vegetation. Some individuals appear to be foraging while **others move across the rocky terrain.**

Figure 4: Behavior-captioning comparisons on chimpanzee and baboon clips. Colored text distinguishes GT-supported details from unsupported or speculative content.