BEYOND ACTIONS AND OBJECTS: VIDEO CLASSIFICATION IN THE WILD

Anonymous authors
Paper under double-blind review

ABSTRACT

Existing video classification datasets are largely built around predefined actions, objects, or their combinations. However, video classification in the wild is more challenging: videos within the same category can vary widely in subjects, scenes, and appearance, while different categories can be visually similar yet semantically distinct. To study this problem, we introduce **EventTree**, a human-in-the-loop framework that discovers and organizes event categories from large-scale video collections through top-down taxonomy construction. Applying EventTree yields **WildEvent-1K**, a hierarchical event-centric video classification dataset with 538K videos and $1,040$ fine-grained event categories in a three-level taxonomy. WildEvent-1K exhibits substantial intra-class variation, subtle semantic differences between neighboring categories, and a long-tailed distribution, providing a challenging testbed for real-world classification. These characteristics make conventional video encoders with direct classification heads less effective, motivating classification through next-token prediction with a VLM pretrained across diverse visual domains. To accommodate the hierarchical structure of WildEvent-1K, we propose **Generative Hierarchical Video Classification (GHVC)**, which predicts categories level by level, restricting each prediction to valid children of its predicted parent. Experiments reveal a clear contrast with conventional datasets such as Kinetics-400: specialized video encoders remain strong in action recognition but are substantially less effective on WildEvent-1K, where VLM-based classification performs better. GHVC further improves it through hierarchical next-token prediction, demonstrating its effectiveness for video classification in the wild. Models, code, and data will be publicly released.

1 INTRODUCTION

Video classification is fundamental to video understanding. Most existing datasets use predefined actions, objects, or combinations, such as "running", "playing guitar", or "riding a horse" (Soomro et al., 2012; Kay et al., 2017; Goyal et al., 2017; Monfort et al., 2019). These datasets have driven substantial progress, yet real-world videos often go beyond such visually distinctive categories, making video classification more challenging, as shown in Figure 1.

Practical applications such as content moderation require event-level understanding beyond recognizing actions or objects. For example, reviewing videos for speeding or drunk-driving-related behavior requires interpreting driving in context. The same event category can include footage captured on highways or residential roads, during the day or at night, using dashboard-mounted or helmet-mounted cameras. Conversely, visually similar driving scenes may depict semantically distinct events. These application-driven requirements motivate our study of **video classification in the wild**: recognizing shared event semantics across diverse appearances while distinguishing visually similar but semantically different events.

To study this problem, we introduce **EventTree**, a human-in-the-loop framework that progressively discovers and organizes event categories from large-scale video collections through top-down taxonomy construction. Rather than defining a complete category vocabulary in advance, EventTree starts from broad groups of heterogeneous videos and recursively refines them into increasingly specific event categories. A VLM generates event summaries and semantic attributes, GPT analyzes these records to propose candidate categories, and humans inspect original videos to verify category

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.

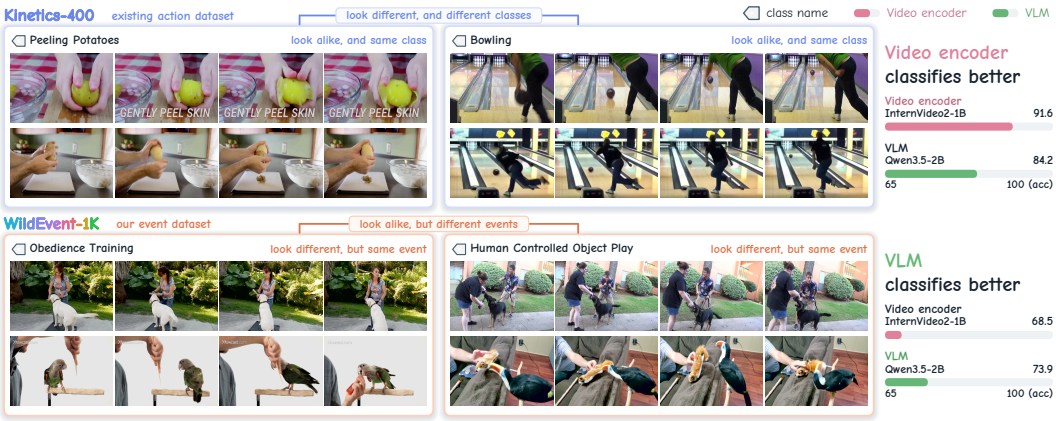


Figure 1: **Motivation for video classification in the wild.** Top: Kinetics-400 action classes are relatively visually consistent and separable, favoring specialized video encoders. Bottom: WildEvent-1K exhibits larger intra-class variation and subtler inter-class differences, where VLM-based classification becomes more effective. Representative examples and model performance illustrate this contrast.

boundaries, resolve ambiguities, and merge or remove unstable categories. Through this iterative process, the taxonomy is constructed from the observed video collection rather than imposed by a predefined action vocabulary.

Applying EventTree to VidGen-1M (Tan et al., 2024) yields **WildEvent-1K**, a hierarchical event-centric video classification dataset with 538K videos and $1,040$ fine-grained event categories (leaf) in a three-level taxonomy, including 12 top-level event families (L1) and 129 intermediate categories (L2). WildEvent-1K exhibits substantial intra-class variation, subtle differences between neighboring categories, and a long-tailed distribution. As shown in Figure 1, videos within the same category can differ greatly in appearance, while different categories can remain visually similar despite distinct event semantics. These properties better reflect many real-world classification scenarios, making WildEvent-1K a challenging testbed and foundation for studying video classification in the wild.

These characteristics make conventional video encoders with direct classification heads less effective. Specialized video encoders perform strongly on action-centric benchmarks such as Kinetics-400 (Kay et al., 2017), where characteristic motion and appearance provide strong classification cues. On WildEvent-1K, however, their performance drops substantially as category distinctions depend more on event semantics. This motivates us to instead perform classification through next-token prediction with a VLM that has been broadly pretrained across diverse visual domains.

To accommodate the hierarchical structure of WildEvent-1K, we further propose **Generative Hierarchical Video Classification (GHVC)**. GHVC represents each taxonomy node with a dedicated category token or category name and predicts categories level by level. Starting from the top level, the model predicts one category at each step, with each subsequent prediction restricted to the valid children of the previously predicted parent until a leaf category is reached. In this way, the hierarchical taxonomy is directly incorporated into VLM classification while retaining next-token prediction.

Experiments reveal a clear contrast between conventional action recognition and video classification in the wild. Specialized video encoders remain strong on Kinetics-400, but are substantially less effective on WildEvent-1K, where VLM-based classification achieves considerably stronger performance. GHVC further improves the VLM-based classification by incorporating the hierarchical taxonomy into next-token prediction, consistently outperforming conventional video encoders on WildEvent-1K.

Our main contributions are summarized as follows.

- We introduce **EventTree**, a human-in-the-loop framework that discovers and organizes event categories from large-scale video collections through top-down taxonomy construction.
- We construct **WildEvent-1K**, a hierarchical event-centric video classification dataset with approximately 538K videos, 12 top-level event families, 129 intermediate categories, and $1,040$ fine-grained event categories. WildEvent-1K features substantial intra-class variation, subtle semantic differences between neighboring categories, and a long-tailed class distribution, providing a challenging testbed for video classification in the wild.

- We show that specialized video encoders strong on conventional action recognition benchmarks are substantially less effective on WildEvent-1K. We further propose **GHVC**, which performs generative hierarchical classification through level-by-level next-token prediction.

2 Related Work

Video Classification Datasets. Video classification datasets have largely focused on human actions. UCF101 (Soomro et al., 2012) and Kinetics (Kay et al., 2017) provide large action vocabularies, while Something-Something (Goyal et al., 2017) emphasizes temporally sensitive human–object interactions. Charades (Sigurdsson et al., 2016) covers multi-label indoor activities, EPIC-KITCHENS-100 (Damen et al., 2022) fine-grained egocentric actions and object interactions, and Moments in Time (Monfort et al., 2019) short dynamic events involving people, objects, animals, and natural phenomena. Richer label structures also appear in ActivityNet (Caba Heilbron et al., 2015), which organizes human activities hierarchically, FineGym (Shao et al., 2020), which provides a three-level gymnastics hierarchy, and PortraitMode-400 (Han et al., 2024), which builds a data-driven taxonomy for portrait videos. YouTube-8M (Abu-El-Haija et al., 2016) and Holistic Video Understanding (Diba et al., 2020) further scale to thousands of actions, objects, scenes, and concepts, mainly for multi-label recognition. In contrast, EventTree constructs a broad-domain, three-level event taxonomy directly from the observed video distribution, assigning each video a coherent root-to-leaf category.

Video Classification Methods. Conventional video classification uses a visual encoder with a task-specific classifier, including Two-Stream Networks (Simonyan & Zisserman, 2014), I3D (Carreira & Zisserman, 2017), SlowFast (Feichtenhofer et al., 2019), TimeSformer (Bertasius et al., 2021), Video Swin Transformer (Liu et al., 2022), and VideoMAE (Tong et al., 2022). Foundation models such as InternVideo (Wang et al., 2022) and InternVideo2 (Wang et al., 2024) scale video pretraining with masked modeling and multimodal supervision, but generally retain task-specific heads for closed-set classification. Video–language methods instead exploit textual class semantics. ActionCLIP (Wang et al., 2021), X-CLIP (Ni et al., 2022), ViFi-CLIP (Rasheed et al., 2023), and Open-VCLIP (Weng et al., 2023) perform recognition through video–text alignment, while recent MLLM-based methods use language-oriented prediction, including multiple-choice classification in LLaVAction (Qi et al., 2026) and discriminative–generative formulations in Pang et al. (2026). Hierarchical video classification further models coarse-to-fine label relations through multi-level supervision or hierarchical constraints (Zhang et al., 2025). Our GHVC instead performs coarse-to-fine VLM generation, predicting a hierarchy-consistent root-to-leaf path through next-token prediction.

3 EventTree: Building WildEvent-1K

3.1 Overview

Given a large collection of unconstrained videos, our goal is to construct a hierarchical taxonomy of *what is happening* without restricting the label space to predefined actions or objects. We introduce **EventTree**, a human-in-the-loop framework that progressively organizes videos into increasingly fine-grained event categories through a top-down process (Figure 2): VLM-generated descriptions are analyzed by GPT to propose coarse-to-fine event candidates, while human verification refines their definitions and boundaries. Applying EventTree to VidGen-1M (Tan et al., 2024) produces **WildEvent-1K**, containing 538,539 videos organized into a three-level hierarchy with 12 top-level event families (L1), 129 intermediate categories (L2), and 1,040 fine-grained event categories (leaf). A detailed overview of the construction pipeline is provided in Appendix A.1.

3.2 Top-Down Category Discovery

EventTree defines categories by the dominant visible event in a video. An event may jointly depend on actions, participants, interaction targets, relevant objects, relations, state changes, and temporal context, but all category distinctions must be supported by observable video evidence.

We use Qwen3-VL-235B (Bai et al., 2025) to generate event summaries and semantic attributes, and GPT-5.6 (OpenAI, 2026) to propose candidate categories from these records for human verification; neither output is treated as ground truth. Global and L1-specific prompting schemes extract

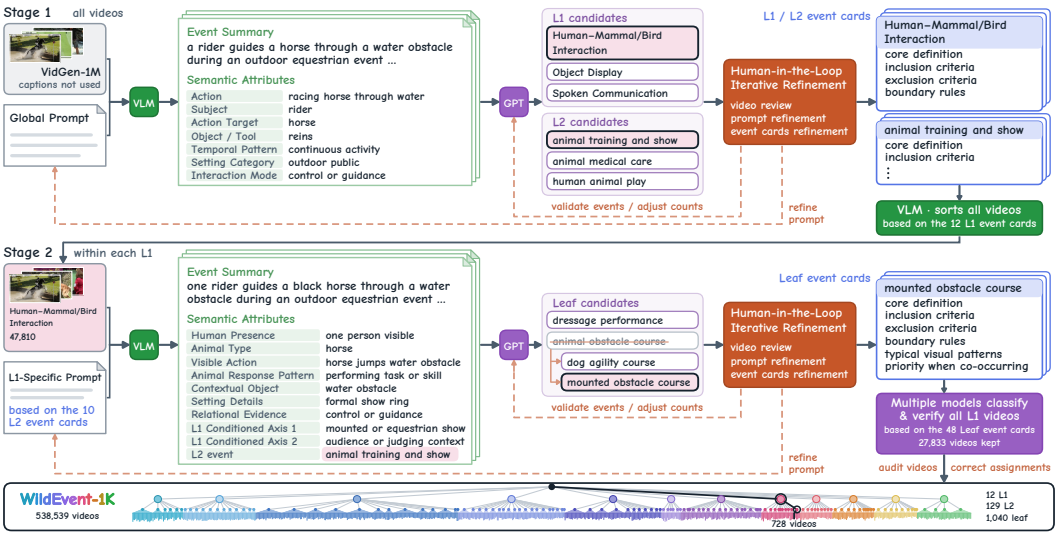
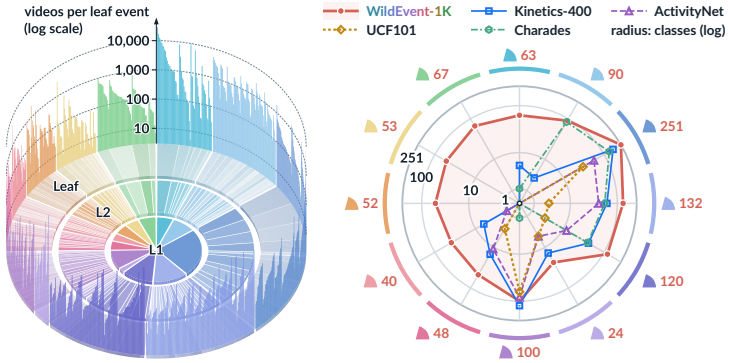


Figure 2: **Overview of the EventTree construction pipeline.** GPT analyzes VLM-generated summaries and semantic attributes to propose L1 and L2 event candidates, while human review validates them and refines prompts and event cards. These cards guide video assignment, and the process repeats within each L1 branch to discover leaf categories. After taxonomy finalization, multi-model classification and verification, followed by human auditing, produce the final WildEvent-1K annotations.



L1	L2	Leaf	videos
Spoken Communication	6	63	164,307
Object Display	12	90	100,400
Life and Society	13	251	63,996
Making, Repair and Operation	19	132	40,797
Food and Beverage	11	120	33,621
Vehicle Movement	6	24	33,193
Sports and Fitness	18	100	26,051
Human–Mammal/ Bird Interaction	10	48	27,833
Human–Aquatic/ Other Interaction	9	40	14,031
Mammals and Birds	11	52	13,306
Aquatic and Other	10	53	7,372
Scenes and Environments	4	67	13,632
All	12	129	1,040	538,539

Figure 3: **WildEvent-1K at a glance. Left**: three-level taxonomy and long-tailed video distribution over $1,040$ leaf categories. **Middle**: category coverage versus prior video classification datasets; WildEvent-1K spans all 12 event families with substantially broader and finer-grained categories within each. **Right**: taxonomy and video-distribution statistics by event family.

complementary semantic evidence for this process, as detailed in Appendix A.2. EventTree first identifies broad recurring event families and then recursively subdivides heterogeneous groups using a local divisive strategy. Candidate splits are retained only when they represent recurring and visually verifiable event distinctions; weak or unstable partitions are merged or rejected. Details of candidate-category construction and event-card specification are provided in Appendix A.3. Each video is ultimately assigned to one dominant root-to-leaf event path.

3.3 Human-in-the-Loop Refinement

Model-discovered categories are treated as candidates rather than final supervision. Human reviewers inspect the original videos to validate candidate events, refine category boundaries and granularity through splitting, merging, removal, or redefinition, and correct ambiguous or incorrectly routed samples.

Recurring ambiguities are further consolidated into reusable category definitions and boundary rules, progressively stabilizing the taxonomy. Representative refinement cases and residual-driven category discovery are provided in Appendices A.3 and A.4.

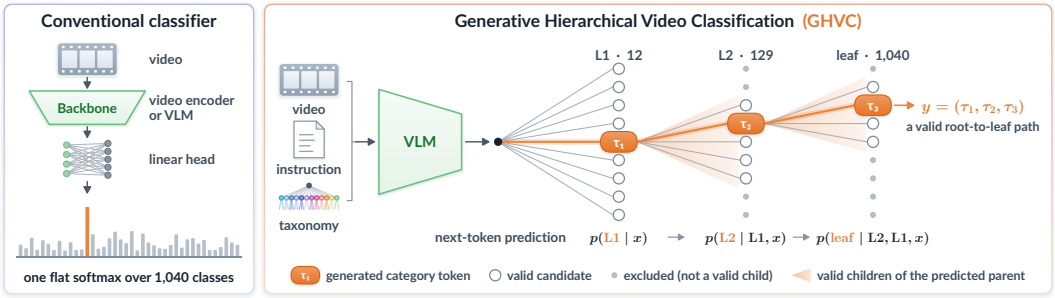


Figure 4: **Overview of Generative Hierarchical Video Classification (GHVC).** Given sampled video frames, a prompt, and the taxonomy, GHVC generates a root-to-leaf path through next-token prediction. In the default implementation, each taxonomy node is represented by a dedicated category token. Training uses teacher forcing with answer-only cross-entropy, while inference restricts each prediction to valid children of the parent, ensuring a hierarchy-consistent path.

3.4 WildEvent-1K

After taxonomy construction, video–label assignments are filtered through multi-model agreement and verification, followed by a category-stratified human audit (Appendices A.5 and A.6). The resulting WildEvent-1K contains 538,539 videos and 1,040 leaf event classes.

WildEvent-1K combines a large video collection with a substantially finer and broader event label space than conventional action-centric benchmarks. Table 6 in Appendix A.7 compares its class scale and long-tailed distribution, while Figure 3 highlights its coverage across diverse event types.

Together, its broad semantic coverage, large intra-class variation, subtle inter-class boundaries, and long-tailed class distribution make WildEvent-1K a distinct setting from conventional action recognition and better reflect many real-world classification challenges. WildEvent-1K therefore provides a foundation for studying video classification methods intended for practical applications.

4 Generative Hierarchical Video Classification

4.1 Motivation and Overview

Conventional video classification relies on video encoders with linear heads, but we find this formulation substantially less effective on WildEvent-1K, which features large intra-class variation, subtle inter-class differences, and broad semantic coverage. We therefore turn to VLMs, whose large-scale multimodal pretraining provides rich semantic knowledge across diverse visual domains, and formulate classification as next-token generation rather than linear-head prediction.

To exploit the hierarchical structure of WildEvent-1K, we propose **Generative Hierarchical Video Classification (GHVC)**, which generates a coarse-to-fine category path, conditioning each prediction on its parent category. Figure 4 illustrates the framework.

4.2 Hierarchical Generation

Output representation. Let the three-level taxonomy be a tree $\mathcal{T} = (\mathcal{N}, \mathcal{E})$. For a video V, its ground-truth path is

$$\pi(V) = (n_1, n_2, n_3), \tag{1}$$

where n_1, n_2, and n_3 denote its L_1, L_2, and leaf categories. GHVC represents this path either by generating the natural-language category names $c(n_1) \to c(n_2) \to c(n_3)$, or by assigning each non-root node n a dedicated category token τ_n and generating

$$\mathbf{y} = (\tau_{n_1}, \tau_{n_2}, \tau_{n_3}). \tag{2}$$

The latter yields 1,181 category tokens for WildEvent-1K and serves as our default implementation, providing a fixed three-token target instead of variable-length category names.

Each category token is initialized from its natural-language label. If the label is tokenized as $(w_1, \ldots, w_K)$,

$$E^{(0)}_{\tau_n} = \frac{1}{K} \sum_{k=1}^{K} E_{w_k}, \tag{3}$$

providing a semantic initialization rather than an arbitrary class symbol.

Autoregressive training. Given sampled video frames, a classification instruction I, and the taxonomy $\mathcal{T}$, GHVC predicts the ground-truth path autoregressively with teacher forcing:

$$\mathcal{L}_{\mathrm{GHVC}} = -\sum_{\ell=1}^{3} \log p_\theta \left(\tau_{n_\ell} \mid V, \mathcal{T}, I, \tau_{n_{<\ell}} \right). \tag{4}$$

Each finer prediction is therefore conditioned on its ancestors, incorporating the hierarchy through the taxonomy context and coarse-to-fine target sequence without an additional hierarchical loss. Only answer tokens contribute to the loss. The category-name representation follows the same procedure using natural-language labels instead of category tokens.

Hierarchy-constrained inference. Autoregressive generation does not by itself guarantee a valid taxonomy path. Let $\mathrm{Ch}(p)$ denote the children of node p. Starting from $\hat{n}_0 = r$, we restrict prediction at level ℓ to

$$\hat{n}_\ell = \arg \max_{n \in \mathrm{Ch}(\hat{n}_{\ell-1})} z^{(\ell)}_{\tau_n}, \tag{5}$$

where r is the root and $z^{(\ell)}_{\tau_n}$ is the logit of τ_n at step ℓ. The selected category is appended to the prefix and prediction continues until a leaf is reached. This reduces each decision to the relevant local branch while guaranteeing a valid root-to-leaf path.

Additional implementation details are provided in Appendix B: category tokens are uniquely associated with taxonomy nodes, and training retains the standard LM formulation without a separate classification head or sibling-specific normalization (Appendices B.1 and B.2). We further introduce an efficient answer-only implementation that preserves the full multimodal and taxonomy context while applying the vocabulary projection only to supervised answer positions (Appendix B.4).

5 Experiments

5.1 Experimental Setup

We conduct primary experiments on WildEvent-1K, reserving 50K videos for testing and using the remainder for training, and additionally evaluate on Kinetics-400 (K400) (Kay et al., 2017). We compare Video Swin Transformer (Liu et al., 2022), VideoMAE (Tong et al., 2022), and InternVideo2 (Wang et al., 2024) with Qwen3.5 (Team, 2026). Unless otherwise stated, VLMs are adapted for two epochs with LoRA while freezing the vision tower, and we evaluate classifier heads, text generation, flat category tokens, and GHVC. For WildEvent-1K, we report Acc_{L1}, Acc_{L2}, Acc_{Leaf}, and leaf macro-F1, with tree distance, hierarchy consistency, parse failure rate, and head/medium/tail F1 when applicable; K400 uses Top-1 accuracy. Additional details are provided in Appendix C.

5.2 Video Encoders versus VLMs

We first disentangle representation choice from classification formulation. With conventional classification heads, Table 1 shows a clear reversal between K400 and WildEvent-1K. On K400 (Kay et al., 2017), InternVideo2-1B (Wang et al., 2024) achieves 91.6% Top-1 accuracy, outperforming Qwen3.5-2B by 7.42%. On WildEvent-1K, frozen probing still favors specialized video encoders, but the trend reverses after downstream adaptation: with only 2 epochs of LoRA training, Qwen3.5-2B reaches 72.83% leaf accuracy and 71.22% macro-F1, already surpassing InternVideo2-1B after 15 epochs of full fine-tuning (68.46%/65.28%). Qwen3.5-9B further reaches 75.02%/74.62%.

GHVC provides consistent gains beyond the VLM backbone itself. Across matched 2B, 4B, and 9B models, GHVC improves the linear-head baseline by 1.09%–1.36% in leaf accuracy and 1.97%–2.22% in macro-F1. Notably, Qwen3.5-4B with GHVC surpasses the 9B linear-head baseline

Table 1: **Specialized video encoders versus VLM-based classification on Kinetics-400 and WildEvent-1K.** K400 reports Top-1 accuracy, while WildEvent-1K reports leaf accuracy and macro-F1. Video encoders lead on K400, whereas adapted VLMs perform better on WildEvent-1K; GHVC further improves the same VLM backbones over linear-head classification.

			Training / Model			Performance	
Paradigm	**Backbone**	**Adaptation**	**Epochs**	**Params**	**Frames**	**Accuracy**	**mF1**
		Kinetics-400					
Video encoder	Video Swin-L (Liu et al., 2022)	Full FT	30	197M	32	83.1%	–
Video encoder	VideoMAE-L (Tong et al., 2022)	Full FT	50	304M	16	86.4%	–
Video encoder	VideoMAE-H (Tong et al., 2022)	Full FT	50	632M	16	86.9%	–
Video encoder	InternVideo2-1B (Wang et al., 2024)	Full FT	3	1B	16	**91.6%**	–
VLM + linear head	Qwen3.5-2B (Team, 2026)	LoRA	2	2B	2 fps (∼15.6)	84.18%	–
		WildEvent-1K					
Frozen backbone: linear probing							
Video encoder	Video Swin-L (Liu et al., 2022)	Linear probe	100	197M	16	54.50%	51.09%
Video encoder	VideoMAE-B (Tong et al., 2022)	Linear probe	100	86M	16	47.41%	40.12%
Video encoder	VideoMAE-L (Tong et al., 2022)	Linear probe	100	304M	16	55.10%	52.75%
Video encoder	VideoMAE-H (Tong et al., 2022)	Linear probe	100	632M	16	57.12%	52.93%
Video encoder	InternVideo2-1B (Wang et al., 2024)	Linear probe	15	1B	16	**63.20%**	**62.65%**
VLM + linear head	Qwen3.5-2B (Team, 2026)	Linear probe	2	2B	2 fps (∼14.3)	34.19%	17.59%
VLM + linear head	Qwen3.5-4B (Team, 2026)	Linear probe	2	4B	2 fps (∼14.3)	54.19%	43.44%
VLM + linear head	Qwen3.5-9B (Team, 2026)	Linear probe	2	9B	2 fps (∼14.3)	50.82%	36.29%
Zero-shot generative classification							
VLM + GHVC	Qwen3.5-2B (Team, 2026)	Zero-shot	–	2B	2 fps (∼14.3)	3.92%	3.93%
VLM + GHVC	Qwen3.5-4B (Team, 2026)	Zero-shot	–	4B	2 fps (∼14.3)	31.24%	32.78%
VLM + GHVC	Qwen3.5-9B (Team, 2026)	Zero-shot	–	9B	2 fps (∼14.3)	**53.84%**	**54.46%**
VLM + GHVC	Qwen3.5-35B-A3B (Team, 2026)	Zero-shot	–	35B	2 fps (∼14.3)	43.67%	45.02%
Downstream adaptation							
Video encoder	Video Swin-L (Liu et al., 2022)	Full FT	30	197M	16	58.33%	53.03%
Video encoder	VideoMAE-B (Tong et al., 2022)	Full FT	30	86M	16	51.35%	42.05%
Video encoder	VideoMAE-L (Tong et al., 2022)	Full FT	30	304M	16	60.01%	53.68%
Video encoder	VideoMAE-H (Tong et al., 2022)	Full FT	30	632M	16	60.24%	52.48%
Video encoder	InternVideo2-1B (Wang et al., 2024)	Full FT	15	1B	16	68.46%	65.28%
VLM + linear head	Qwen3.5-2B (Team, 2026)	LoRA	2	2B	2 fps (∼14.3)	72.83%	71.22%
VLM + linear head	Qwen3.5-4B (Team, 2026)	LoRA	2	4B	2 fps (∼14.3)	74.17%	73.18%
VLM + linear head	Qwen3.5-9B (Team, 2026)	LoRA	2	9B	2 fps (∼14.3)	75.02%	74.62%
VLM + GHVC	Qwen3.5-2B (Team, 2026)	LoRA	2	2B	2 fps (∼14.3)	73.92%	73.19%
VLM + GHVC	Qwen3.5-4B (Team, 2026)	LoRA	2	4B	2 fps (∼14.3)	75.53%	75.40%
VLM + GHVC	Qwen3.5-9B (Team, 2026)	LoRA	2	9B	2 fps (∼14.3)	76.37%	76.75%
VLM + GHVC	Qwen3.5-35B-A3B (Team, 2026)	LoRA	2	35B	2 fps (∼14.3)	**76.87%**	**77.47%**

(75.53%/75.40% vs. 75.02%/74.62%), indicating that the improvement is not simply due to a larger VLM backbone. Scaling GHVC to Qwen3.5-35B-A3B further reaches 76.87% leaf accuracy and 77.47% macro-F1.

5.3 Analysis of Classification Formulations

We next isolate the effect of the classification formulation using Qwen3.5-2B under the same training protocol. Table 2 compares conventional linear-head classification, flat category-name generation, flat category-token generation, and GHVC.

Replacing the linear classifier with generative prediction already improves performance: flat name generation increases leaf accuracy from 72.83% to 73.74%, while a single category token reaches 73.78%. Their similar performance indicates that the gain mainly comes from generative classification rather than the token representation itself. GHVC further improves all hierarchy levels, achieving 73.92% leaf accuracy and 73.19% macro-F1, or gains of 1.09% and 1.97% over the linear-head baseline. Its consistent improvement over both flat formulations shows that explicitly following the taxonomy provides additional benefit beyond VLM generation alone.

Inference efficiency. Table 3 compares the decoding spaces of different generative formulations. Flat name generation predicts variable-length text over the full language vocabulary, while a flat category token requires a single global decision among $1,040$ leaf classes. GHVC instead decomposes classification into three small hierarchical decisions, reducing the candidate space to the local children

Table 2: **Ablation of VLM classification formulations on WildEvent-1K using Qwen3.5-2B.** All methods use the same training protocol. "Gen." denotes next-token generation, "Sem." category semantics, and "Hier." hierarchical prediction. Moving from a linear head to generative classification improves performance, while GHVC further benefits from coarse-to-fine taxonomy prediction.

	Classification Formulation				Recognition				Hierarchy		
Method	**Gen.**	**Sem.**	**Hier.**	**Output**	Acc_{L1}	Acc_{L2}	Acc_{Leaf}	**Leaf mF1**	**TreeDist ↓**	**Cons. ↑**	**Parse fail ↓**
VLM + Linear Head	✗	✗	✗	Class index	94.94%	84.44%	72.83%	71.22%	0.95	100.0	0.0
Flat Name Generation	✓	✓	✗	Label string	95.19%	84.83%	73.74%	72.94%	0.92	99.94	0.0
Flat Category Token	✓	✓	✗	Leaf token	94.76%	84.70%	73.78%	72.86%	0.93	100.0	0.0
GHVC	✓	✓	✓	Category path	**95.28%**	**84.97%**	**73.92%**	**73.19%**	**0.92**	**100.0**	**0.0**

Table 3: **Accuracy–efficiency comparison of generative classification formulations.** GHVC replaces a global leaf decision with coarse-to-fine decisions, achieving best accuracy at comparable decoding efficiency. Throughput is measured in videos/s on 8× H100 GPUs at batch size 16.

	Decoding			Performance	
Method	**Output tokens**	**Candidate space**	**Throughput ↑**	**Leaf Acc. ↑**	**mF1 ↑**
VLM + linear head	–	1040	36.0	72.83%	71.22%
Flat Name Generation	Variable	$\lvert\mathcal{V}\rvert$ per token	21.3	73.74%	72.94%
Flat Category Token	1	1,040	35.1	73.78%	72.86%
GHVC	3	$12 \rightarrow \sim 10.8 \rightarrow \sim 8.1$	34.1	**73.92%**	**73.19%**

of each predicted parent. This structured decoding adds little latency over a flat category token while avoiding a global fine-grained decisio

Long-tail behavior. Figure 5 shows the accuracy and mF1 of GHVC across classes with different training frequencies. On head and medium classes, the mean per-class accuracy gain is only $+0.2\%$, $+1.0\%$, and $+1.1\%$ for the 2B, 4B, and 9B models, but rises to $+7.8\%$, $+9.5\%$, and $+9.4\%$ on tail classes. The same pattern holds for F1, with tail gains of $+6.9\%$, $+7.9\%$, and $+7.3\%$. The gap further widens toward the rarest classes and remains large as the backbone scales, showing that GHVC is particularly effective in the data-scarce regime.

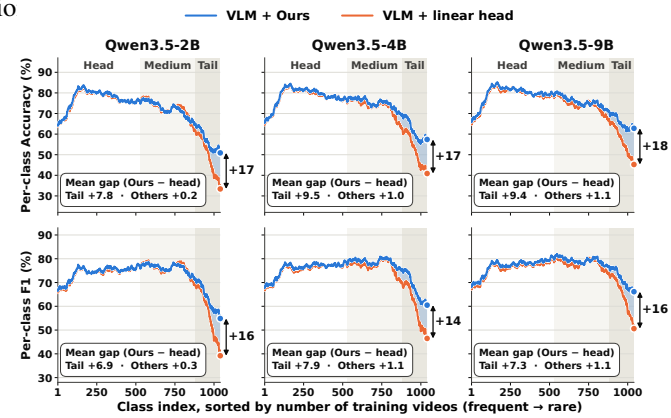


Figure 5: **Per-class accuracy and F1 across the WildEvent-1K label space.** Classes are sorted from most to least frequent in training set and smoothed with an 80-class moving average.

Why does GHVC help rare classes? The linear head exhibits a strong frequency bias: rare classes are often under-predicted and absorbed by more frequent categories. GHVC instead predicts through the VLM output space, allowing rare categories to benefit from pretrained semantic knowledge rather than learning each class solely from a few examples. Consistently, post-hoc logit adjustment raises the 2B linear head's tail macro-F1 from 56.6% to 65.4%, but reduces head accuracy by 1.4%. GHVC reaches 63.6% tail macro-F1 without class-frequency priors or tuned correction, suggesting that its advantage largely comes from avoiding the frequency bias introduced by a conventional linear head.

5.4 Scaling and Adaptation of GHVC

We next study how model scale and adaptation strategy affect GHVC. Table 4 shows that zero-shot recognition benefits strongly from model scale: leaf accuracy rises from 3.92% at 2B to 31.24% at 4B and 53.84% at 9B, indicating substantially stronger event-semantic knowledge in larger pretrained VLMs. After only two epochs of adaptation with GHVC, this gap narrows markedly: LoRA reaches 73.92%, 75.53%, and 76.37% leaf accuracy at 2B, 4B, and 9B, respectively, showing diminishing returns from model scale after supervised adaptation.

LoRA consistently outperforms full fine-tuning while updating only 0.19%–0.33% of the parameters. At 9B, it improves leaf accuracy and macro-F1 by 2.17% and 2.94%, with an 8.05% gain on tail F1. Notably, the 4B LoRA model surpasses the 9B full-finetuned model (75.53%/75.40% vs.

Table 4: **Scaling and adaptation of GHVC on WildEvent-1K.** Zero-shot performance improves strongly with model scale, while supervised adaptation substantially narrows the gap.

		Trainable	Overall				Frequency Groups		
Backbone	**Adaptation**	**Params**	**Acc$_{L1}$**	**Acc$_{L2}$**	**Acc$_{Leaf}$**	**Leaf mF1**	**Head F1**	**Medium F1**	**Tail F1**
Qwen3.5-2B	Zero-shot	0	8.55%	5.30%	3.92%	3.93%	4.11%	2.90%	5.86%
	LoRA	7.38M	**95.28%**	**84.97%**	**73.92%**	**73.19%**	**74.11%**	**76.09%**	**63.56%**
	Full FT	2B	94.83%	84.42%	73.37%	72.32%	73.81%	75.35%	60.51%
Qwen3.5-4B	Zero-shot	0	47.27%	39.07%	31.24%	32.78%	30.02%	35.92%	34.29%
	LoRA	14.38M	**95.68%**	**86.07%**	**75.53%**	**75.40%**	**75.64%**	**78.65%**	**67.69%**
	Full FT	4B	94.88%	84.56%	73.51%	72.71%	74.04%	75.29%	62.48%
Qwen3.5-9B	Zero-shot	0	80.94%	67.40%	53.84%	54.46%	51.78%	58.28%	54.26%
	LoRA	18.11M	**95.91%**	**86.51%**	**76.37%**	**76.75%**	**76.63%**	**79.20%**	**71.82%**
	Full FT	9B	95.07%	84.96%	74.20%	73.81%	74.67%	76.99%	63.77%

74.20%/73.81% in leaf accuracy/macro-F1) while training only 14.38M parameters. These results show that GHVC supports efficient adaptation, with particularly strong gains on long-tailed categories.

5.5 Analysis of Class Geometry

We assess whether WildEvent-1K exhibits greater intra-class variation and subtler inter-class boundaries than K400. Using frozen InternVideo2-1B (Wang et al., 2024) and X-CLIP-B/16 (Ni et al., 2022), we compute intra-class cosine distance and nearest-class centroid similarity over 50 matched draws of 400 classes, with up to 20 videos per class.

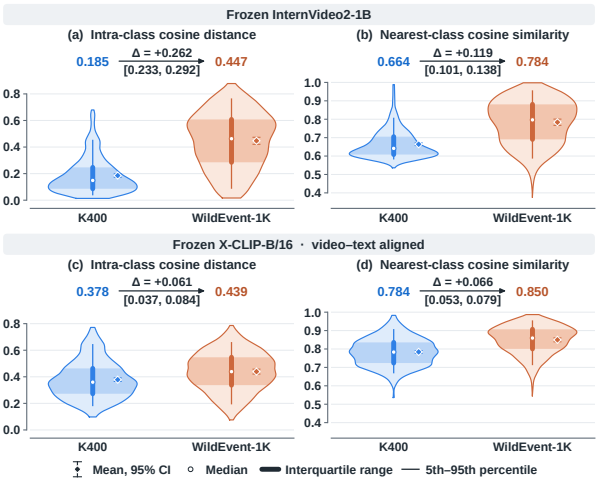


Figure 6: **Class geometry on K400 and WildEvent-1K.** WildEvent-1K shows larger intra-class variation and higher nearest-class similarity under frozen InternVideo2-1B and X-CLIP (Ni et al., 2022).

Figure 6 shows consistently harder class geometry in WildEvent-1K. With InternVideo2-1B, intra-class distance increases from 0.185 to 0.447 (+0.262), while nearest-class similarity rises from 0.664 to 0.784 (+0.120). X-CLIP-B/16 shows the same trend, increasing from 0.378 to 0.439 (+0.061) and from 0.784 to 0.850 (+0.066), respectively. The consistent effects across matched draws indicate greater intra-class variation and subtler inter-class boundaries in WildEvent-1K across both representation spaces.

6 Conclusion

In this work, we study video classification in the wild, characterized by broad semantic coverage, large intra-class variation, subtle inter-class differences, and long-tailed distributions. We introduce **EventTree**, a human-in-the-loop framework for constructing hierarchical event taxonomies from large-scale video collections, yielding **WildEvent-1K** with 538K videos and 1,040 fine-grained event categories. Our experiments reveal a shift from conventional action recognition: specialized video encoders remain strong on Kinetics-400, while pretrained VLMs perform better on WildEvent-1K after downstream adaptation. We further propose **Generative Hierarchical Video Classification (GHVC)**, which performs coarse-to-fine next-token prediction along the taxonomy and consistently improves VLM-based classification. Together, WildEvent-1K and GHVC provide a challenging benchmark and effective framework for structured event classification in real-world videos.

Limitations. 1) Human involvement. EventTree requires significant human effort, limiting full automation and scalability. **2) Single-path labels.** Each video is assigned one dominant root-to-leaf path, which may simplify videos with multiple concurrent or sequential events. **3) Evaluation scope.** WildEvent-1K is constructed from VidGen-1M, and GHVC is evaluated mainly with Qwen3.5 models; broader validation across data sources and VLM families remains future work.

AI USE STATEMENT

AI models were used in dataset construction, as described in the main text. Otherwise, generative AI tools were used only for language polishing, minor consistency checks, and improving the presentation of figures. All research decisions, results, claims, figures, and final content were reviewed and verified by the authors.

REPRODUCIBILITY STATEMENT

Appendix A documents dataset construction, prompt and output-field design, event-card specifications, and human auditing. Appendix B details the GHVC implementation, while Appendix C specifies the data splits, preprocessing, training configurations, and evaluation protocols. The complete set of L1, L2, and leaf event cards is provided as supplementary material. Models, code, and data will be publicly released.

ETHICS STATEMENT

WildEvent-1K is derived from VidGen-1M for research on observable event semantics. Its annotations are not intended as judgments about individuals' identities, hidden intentions, or policy violations. Model-assisted construction and data selection may introduce annotation and coverage biases despite human refinement and auditing. Benchmark performance alone does not establish suitability for automated content moderation or other consequential decisions.

REFERENCES

Sami Abu-El-Haija, Nisarg Kothari, Joonseok Lee, Paul Natsev, George Toderici, Balakrishnan Varadarajan, and Sudheendra Vijayanarasimhan. Youtube-8m: A large-scale video classification benchmark. *arXiv preprint arXiv:1609.08675*, 2016.

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.

Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? *arXiv preprint arXiv:2102.05095*, 2021.

Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the ieee conference on computer vision and pattern recognition*, pp. 961–970, 2015.

Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *2017 IEEE conference on computer vision and pattern recognition (CVPR)*, pp. 4724–4733. IEEE, 2017.

Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Antonino Furnari, Evangelos Kazakos, Jian Ma, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. Rescaling egocentric vision: Collection, pipeline and challenges for epic-kitchens-100. *International Journal of Computer Vision*, 130(1):33–55, 2022.

Ali Diba, Mohsen Fayyaz, Vivek Sharma, Manohar Paluri, Jürgen Gall, Rainer Stiefelhagen, and Luc Van Gool. Large scale holistic video understanding. In *European Conference on Computer Vision*, pp. 593–610. Springer, 2020.

Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *2019 IEEE/CVF international conference on computer vision (ICCV)*, pp. 6201–6210. IEEE, 2019.

Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, et al. The "something something" video database for learning and evaluating visual common sense. In *2017 IEEE international conference on computer vision (ICCV)*, pp. 5843–5851. IEEE, 2017.

Mingfei Han, Linjie Yang, Xiaojie Jin, Jiashi Feng, Xiaojun Chang, and Heng Wang. Video recognition in portrait mode. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21831–21841. IEEE, 2024.

Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017.

Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 3192–3201. IEEE, 2022.

Mathew Monfort, Alex Andonian, Bolei Zhou, Kandan Ramakrishnan, Sarah Adel Bargal, Tom Yan, Lisa Brown, Quanfu Fan, Dan Gutfreund, Carl Vondrick, et al. Moments in time dataset: one million videos for event understanding. *IEEE transactions on pattern analysis and machine intelligence*, 42(2):502–508, 2019.

Bolin Ni, Houwen Peng, Minghao Chen, Songyang Zhang, Gaofeng Meng, Jianlong Fu, Shiming Xiang, and Haibin Ling. Expanding language-image pretrained models for general video recognition. In *European conference on computer vision*, pp. 1–18. Springer, 2022.

OpenAI. GPT-5.6: Frontier intelligence that scales with your ambition, July 2026. OpenAI product release, July 9, 2026.

Zhanzhong Pang, Dibyadip Chatterjee, Fadime Sener, and Angela Yao. On discriminative vs. generative classifiers: Rethinking mllms for action understanding. *arXiv preprint arXiv:2603.02546*, 2026.

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.

Hanoona Rasheed, Muhammad Uzair Khattak, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. Fine-tuned clip models are efficient video learners. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6545–6554. IEEE, 2023.

Dian Shao, Yue Zhao, Bo Dai, and Dahua Lin. Finegym: A hierarchical video dataset for fine-grained action understanding. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2613–2622. IEEE, 2020.

Gunnar A Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *European conference on computer vision*, pp. 510–526. Springer, 2016.

Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. *Advances in neural information processing systems*, 27, 2014.

Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012.

Zhiyu Tan, Xiaomeng Yang, Luozheng Qin, and Hao Li. Vidgen-1m: A large-scale dataset for text-to-video generation. *arXiv preprint arXiv:2408.02629*, 2024.

Gemma Team, Sherif El Abd, Vaibhav Aggarwal, Robin Algayres, Alek Andreev, Olivier Bachem, Ian Ballantyne, Cormac Brick, Victor Cărbune, Michelle Casbon, et al. Gemma 4 technical report. *arXiv preprint arXiv:2607.02770*, 2026.

Qwen Team. Qwen3. 5: Towards native multimodal agents. *URL: https://qwen. ai/blog*, 2026.

Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *Advances in neural information processing systems*, 35:10078–10093, 2022.

Mengmeng Wang, Jiazheng Xing, and Yong Liu. Actionclip: A new paradigm for video action recognition. *arXiv preprint arXiv:2109.08472*, 2021.

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.

Yi Wang, Kunchang Li, Xinhao Li, Jiashuo Yu, Yinan He, Guo Chen, Baoqi Pei, Rongkun Zheng, Zun Wang, Yansong Shi, et al. Internvideo2: Scaling foundation models for multimodal video understanding. In *European conference on computer vision*, pp. 396–416. Springer, 2024.

Zejia Weng, Xitong Yang, Ang Li, Zuxuan Wu, and Yu-Gang Jiang. Open-vclip: Transforming clip to an open-vocabulary video model via interpolated weight optimization. In *International conference on machine learning*, pp. 36978–36989. PMLR, 2023.

Rui Zhang, Shuailong Li, Junxiao Xue, Diying Yan, Feng Lin, Qing Zhang, Xiao Ma, and Xiaoran Yan. Hierarchical action recognition: A contrastive video-language approach with hierarchical interactions. In *Computer Graphics International Conference*, pp. 186–198. Springer, 2025.

A Additional Details of EventTree Construction

This appendix expands the construction pipeline in Figure 2, from video descriptions and event-category discovery to human refinement, model-based label assignment, and final annotation auditing.

A.1 End-to-End Construction Overview

EventTree separates the discovery of an event vocabulary from the assignment of videos to that vocabulary. In the global stage, Qwen3-VL-235B-A22B-Instruct (Bai et al., 2025) generates event summaries and general semantic attributes from videos. GPT (OpenAI, 2026) analyzes these records under the event-category construction principles to propose broad L1 event families and L2 categories. Human inspection of the original videos guides revisions to the candidate definitions and the global prompt; the accepted L1 event cards then guide routing of the video collection.

Within each L1 family, a dedicated local prompt incorporates its L2 definitions and requests attributes suited to the distinctions within that branch. The VLM observes the videos again, without receiving its previous global outputs. GPT uses these L1-specific records to propose finer event categories and their supporting video groups, which are inspected and revised by humans. The global and local stages refer to two rounds of video description, not to the depth of the taxonomy: local records can support repeated subdivision before the final L1–L2–leaf hierarchy is fixed.

Prompt refinement and category refinement are interleaved: a misleading description may require a prompt revision, whereas a coherent description that does not fit the current categories may require a revised category boundary or a new candidate. Once the taxonomy and event cards are finalized, two category-assignment branches and a subsequent verification stage filter candidate labels, followed by a human audit covering every leaf category.

A.2 Global and L1-Specific Semantic Evidence Extraction

Global prompt. The global prompt preserves a compact, cross-domain account of the visible event rather than attempting fine-grained leaf classification. Its event tuples record actions, participants, targets, tools, context, setting, and temporal and interaction structure, with distinct roles for these fields (Table 5(a)). The event summary accompanies these attributes, and `L1 event` records the L1 routing decision only after the descriptive evidence has been produced.

L1-specific prompts. We use 12 local prompts, one for each L1 event family, to capture distinctions that the global schema does not express in sufficient detail. For example, human–animal interactions require separate descriptions of the animal's response and its relation to the person, whereas making and repair require the target, tool, and operation process. Table 5(b) summarizes the information emphasized in each family. The local `L2 event` is a tentative L2 assignment, and L2-conditioned

axes describe distinctions within that proposed L2 event; neither the attributes nor their combinations automatically define new taxonomy levels or leaf categories.

Summary and structured fields. The summary provides a concise, readable account of the event for human inspection, while the structured fields systematically cover a more comprehensive set of semantic dimensions for analysis. Closed-set values normalize recurring patterns, and open short labels retain concrete details, making the structured record particularly useful for cross-video comparison and GPT-assisted discovery of leaf categories. The summary preserves temporal and relational context and remains part of that analysis; discrepancies between descriptions and fields are checked against the original video.

Prompt refinement. We repeatedly compared outputs with sampled original videos across the current categories, recording omitted evidence, unsupported actions or objects, and incorrect routes. Revisions clarified field roles, removed redundant or inference-prone fields, and updated output rules, boundary rules, and examples. In particular, we refined local schemas and output instructions together to make domain-specific actions, relations, and contextual objects explicit. For instance, global tuples describe actions before assigning neutral subjects, while a common local design writes the refined summary after a small participant-presence check and before abstract attributes and classification. These choices aim to limit premature role or category assumptions. Summary instructions prioritize event-relevant actions and relations over decorative background details; open fields allow absent or unidentified objects rather than requiring speculative completion. Previously identified errors were revisited after revisions. Videos whose descriptions or assignments remained unreliable were excluded from the final labeled set.

Example of prompt refinement. The human–dog example illustrates how iterative refinement of local output fields and generation rules makes event-specific evidence explicit. The global output reports `petting_and_shaking_paw`, `none` for `context_object`, and `direct_physical_contact`; its summary also mentions a leather couch. The final local output instead describes the dog first looking at the treat wrapper and then placing its paw on the person's outstretched arm. It records `treat_wrapper`, `control_or_guidance`, and `hand_signal` as separate contextual and relational attributes. These local fields expose the context and cue–response relation beyond generic contact, supporting the distinction between *Obedience Training* and ordinary petting or incidental paw contact.

A.3 Event-Category Discovery and Human Refinement

Event definition and construction principles. An event category captures the dominant observable activity, relation, process, or state that organizes a video. Depending on the domain, its definition may depend on participants, actions, interaction targets, tools, spatial relations, state changes, or temporal context; these are not mandatory slots for every category. In particular, object-display and scene categories need not involve pronounced motion. A candidate must have supporting videos, a coherent shared meaning, and an observable boundary from neighboring categories. An object, setting, or participant relation may be decisive when it is part of the event definition. Low frequency alone does not invalidate a category, and high frequency alone does not justify an incoherent one.

Semantic coherence and visual variation. During refinement, we retain differences in subjects, scenes, viewpoints, and appearance when the defining event semantics remain unchanged, rather than splitting a category to make it visually homogeneous. Conversely, visually similar videos are separated when observable relations or processes support distinct event definitions.

From semantic records to candidate categories. GPT analyzes summaries together with the relevant attributes within a current L1 family or L2 group, identifies recurring event structures, and proposes candidate names, definitions, and supporting video identifiers. The analysis proceeds locally: heterogeneous candidates may be subdivided further, while already coherent groups need not be split for symmetry. Field values and L2-conditioned axes suggest possible distinctions, but their combinations do not automatically become classes. We also re-examine incorrectly routed and unclassified records rather than treating the VLM's proposed L2 event as immutable. This is model-assisted semantic grouping with human decisions, not a fixed-distance clustering algorithm.

Table 5: **Output-field functions and L1-specific information for event discovery.** (a) Representative fields and their distinct roles. (b) Information emphasized within each L1 family. Panel (b) summarizes information dimensions rather than an exhaustive list of output keys; complete schemas are supplied with the prompts.

(a) Functional roles of global and representative local fields

Field(s)	Function and distinction
`action / visible action`; `subject`; `action target`	Describe the visible action, who performs it, and whom or what it acts on, without inferring a professional role.
`object / tool`; `contextual object`; `salient_object_family`	Separate the action's tool or medium from an additional contextual object; the object-family field provides a coarse grouping rather than another object.
`animal type`; `animal_count_scope`; `human presence`	Establish animal identity, number, and visible human presence before interpreting the interaction.
`scene_detail / setting details`; `setting category`	Preserve specific visible scene information separately from a normalized, coarse setting category.
`temporal pattern`; `interaction mode`	Distinguish temporal organization, such as repetition or successive stages, from the broad mechanism of interaction.
`animal response pattern`; `relational evidence`	Record the animal's response pattern separately from the human–animal relation; these abstractions complement the literal action.
`L1 event`; `L2 event`	Record tentative L1 and L2 assignments for organizing analysis, not independent evidence that a category is correct.
`L2-conditioned axis 1`; `L2-conditioned axis 2`	Expose distinctions relevant to the proposed L2 event, providing candidate subdivision cues rather than final leaf labels.
`visible_event_summary`; `refined_event_summary`	Provide a concise event account for human inspection, retaining key relations and temporal context.

(b) Representative differences across the 12 local prompts

L1 event family	Local information and its role in category discovery
Spoken Communication	Speaker configuration, visual anchor, and communication structure distinguish monologues, explanations, interviews, and dialogue.
Object Display	Object identity, presentation format, and state separate display, reveal, rotation, and comparison from active use.
Life and Society	Participant configuration, action targets, and social engagement distinguish social interaction, caregiving, personal activity, and performance.
Making, Repair and Operation	Target, tool, operation mode, and state change reveal the process rather than merely the presence of equipment.
Food and Beverage	Food identity and state, tools, and processing actions distinguish preparation, consumption, serving, and food display.
Vehicle Movement	Vehicle identity, motion, human–vehicle relation, and control evidence distinguish travel from static display or interface operation.
Sports and Fitness	Participant configuration, equipment, action targets, and activity phase distinguish competitive, training, and exercise structures.
Human–Mammal/Bird Interaction	Animal response, contact, guidance, and contextual objects distinguish affection, training, feeding, play, and handling.
Human–Aquatic/Other Interaction	Animal state, handling method, and activity stage distinguish fishing, capture, transfer, hive work, and care.
Mammals and Birds	Animal actions, social relations, and life stage distinguish individual behavior, group interaction, and parental activity.
Aquatic and Other	Locomotion, group structure, feeding, and life-cycle state distinguish movement, collective behavior, predation, and developmental events.
Scenes and Environments	Spatial layout, scene state, environmental dynamics, and viewpoint characterize the environment as the event center.

Human validation and revision. Reviewers inspect samples from every candidate category to check both the fidelity of the descriptions and the proposed common event. They split mixed candidates, merge distinctions that cannot be supported reliably, revise overly restrictive or overlapping definitions, and remove unsupported categories. Misrouted videos are reassigned when another definition is supported. For multi-event videos, the dominant path is selected using the sustained activity, participant relationships, temporal structure, and applicable boundary rules and co-occurrence priorities; a more specific interpretation takes precedence only when its own inclusion criteria are met. Secondary events remain in the descriptions without creating additional labels. Human feedback is recorded in revised cards and, when the problem concerns evidence extraction, fed back into the prompts.

Event cards. Each finalized leaf event card contains six components: Core Definition, Inclusion Criteria, Exclusion Criteria, Boundary Rules, Typical Visual Patterns, and Priority When Co-occurring. Boundary Rules distinguish neighboring events; Priority When Co-occurring separately specifies which event takes precedence when multiple events are present. Evidence may be explicit or established compositionally from the described actions and relations; matching a particular keyword is not required. An omitted nonessential detail is not negative evidence, but a missing defining relation cannot be supplied by inference. Typical visual patterns illustrate admissible instances rather than adding mandatory appearance constraints. The same category meanings guide subsequent model classification and human auditing. The following abridged card illustrates these roles.

> **Leaf event: Obedience Training**. **Taxonomy path:** Human–Mammal/Bird Interaction → Animal Training and Show → obedience training.
>
> **Core Definition.** A person gives a visible cue, gesture, or target and an animal performs a learned, obstacle-free response, such as sitting, staying, giving a paw, or completing a high-five. Food may follow as a reward.
>
> **Inclusion Criteria.** An explicit account of cue-based training or a supported cue–response sequence suffices. Paw contact, a sitting posture, or nearby food without that relation is insufficient.
>
> **Exclusion Criteria.** Direct feeding without a task, free play, physical obstacle-course completion, and ordinary leading or conformation judging are distinguished from this category.
>
> **Boundary Rules.** A learned response to a cue distinguishes training from incidental contact or free play; a task followed by a reward distinguishes it from direct food provision.
>
> **Typical Visual Patterns.** "After a hand cue, the animal raises its paw and completes a high-five response."
>
> **Priority When Co-occurring.** A supported cue–response–reward sequence takes priority over incidental feeding, petting, or paw contact; food or posture alone does not establish training.

Examples of category refinement. Figure 7 shows three mixed candidates corrected by human review. *Horse Show* was separated into no-obstacle precision riding, mounted obstacle completion, and mounted racing; *Dog Show* into agility-course completion and conformation judging. *Instrumental Performance* was separated according to formal ensemble performance without central singing, parade-style instrumental performance, and a coordinated live band with central vocals. These distinctions concern the event structure, not merely the presence of horses, dogs, or musical instruments. The broad names in the figure denote pre-refinement candidates, not additional final leaf categories.

A.4 Residual Analysis and Taxonomy Expansion

During construction, videos not reliably covered by the current cards are retained for residual analysis rather than forced into the nearest leaf. We distinguish incorrect routing, an existing category overlooked by the assignment, insufficient visual evidence, and a coherent event missing from the candidate vocabulary. The first two call for reassignment; the last can motivate category discovery, whereas insufficient evidence does not by itself establish novelty.

Human review searches these records for recurring, visually supported event structures. New candidates undergo the same definition, boundary comparison, and video inspection as other categories. Figure 8 shows six leaf categories recovered through this process, including *Cross-Species Affiliative Interaction* and *Food Package Opening Attempt*. The former requires an affiliative relation beyond



Figure 7: **Mixed categories separated during human-in-the-loop refinement.** Each box is one candidate category that human review found to mix several distinct events, shown together with the leaf categories it was split into. Every leaf is shown with four of its videos, keyed (a)–(d), each sampled as four frames in time order.

mere co-presence; the latter requires a packaging-directed opening attempt rather than food or packaging simply appearing nearby. Residual analysis thus expands coverage without relaxing the evidence required for membership. After the taxonomy is fixed, unresolved videos remain outside the released labeled benchmark.

A.5 Multi-Model Label Assignment and Verification

After finalizing the taxonomy and event cards, we use two category-assignment branches followed by candidate-conditioned verification. All models follow the established classification principles and the card components described above. "High confidence" is a card-based acceptance rule, not a calibrated numerical score: the available evidence must support the core definition, satisfy the inclusion criteria, and respect the applicable boundary rules and co-occurrence priorities, without triggering any exclusion criteria. Insufficient evidence leads to rejection rather than forced assignment.



Figure 8: **Leaf categories recovered from residual samples by human review.** Six WildEvent-1K leaf categories that were absent from the model-proposed candidate set and were introduced only after residual analysis and human video inspection. Each row is one category, shown with four of its videos, keyed (a)–(d), each sampled as four frames in time order.

Description-based assignment. GPT-5.6 (OpenAI, 2026) assigns a leaf category from the L1-specific summaries and semantic attributes previously generated by Qwen3-VL-235B-A22B-Instruct, together with the finalized cards. GPT does not inspect the original video.

Direct video-based assignment. In a separate branch, Qwen3-VL-235B-A22B-Instruct assigns a leaf category directly from the video and finalized cards, without receiving GPT's prediction or its own earlier descriptions. Only videos for which both branches produce the same high-confidence leaf assignment proceed to verification.

Candidate-conditioned verification. Gemma4-31B-it (Team et al., 2026) receives each retained video, its agreed candidate leaf, and the finalized event cards, and makes a binary membership decision. It confirms the assignment only when the high-confidence criteria are met; an incompatible or insufficiently supported assignment is rejected. Only pairs accepted by both assignment branches and confirmed by Gemma enter the model-filtered set for human auditing.

This procedure provides complementary checks rather than three statistically independent annotations: GPT relies on Qwen-generated evidence, and Gemma verifies a proposed label rather than performing unconstrained classification. Human adjudication can subsequently correct or remove labels that passed these checks.

A.6 Human Audit and Annotation Quality Assurance

We conducted a category-stratified audit of **22,500 distinct videos** covering **all 1,040 leaf categories**. Reviewers watched the original videos and consulted the finalized event cards rather than relying on model descriptions. Categories containing at most 50 videos were reviewed in full; those containing

51–5,000 videos contributed 20 audit videos, and those containing more than 5,000 contributed 30. Audit entries were deduplicated by video identifier, and the selected videos originated from distinct source videos.

Three reviewers independently examined every audit video; the set was not divided among reviewers. For a category c assessed by partial sampling (more than 50 assigned videos), let n_c be its number of audited videos and $e_{c,r}$ the number judged incorrectly labeled by reviewer r. Such a category passed only if

$$\frac{e_{c,r}}{n_c} \leq 0.15, \qquad \forall r \in \{1, 2, 3\}.$$

This criterion applies separately to each reviewer, not to a majority vote or an averaged error rate. It allows at most three errors per reviewer in a 20-video sample and four in a 30-video sample. The 15% tolerance applies only to partially sampled categories, not to those reviewed in full.

Reviewers recorded the identifiers of flagged videos and jointly adjudicated them after independent review. A confirmed labeling error was corrected by reassignment only when the video satisfied the destination category's finalized card; otherwise, the video was discarded. A category's passing status did not exempt identified errors from correction. For categories with at most 50 videos, every video was manually examined, and the correction process left no identified labeling errors among the retained videos. All 1,040 categories met the applicable review requirements. This establishes audit coverage across the taxonomy, not exhaustive manual verification of every released video; the 15% threshold for partially sampled categories is a sample-based acceptance rule rather than an estimate of the dataset-wide labeling error rate.

A.7 Dataset Finalization and Supplementary Resources

WildEvent-1K is a curated subset of the one-million-video VidGen-1M collection, rather than an exhaustive annotation of all source videos. We exclude animation, gameplay, and other non-real-world content, which together account for approximately 10.67% of the source collection. Further selection occurs during taxonomy construction: isolated instances lacking support for a stable event category are not incorporated into the developing taxonomy, and human review removes candidate categories that remain semantically heterogeneous or lack a clear event definition. Videos without a supported leaf assignment after refinement and residual analysis are excluded rather than forced into a category; low frequency alone is not an exclusion criterion. Model-based filtering and human auditing further refine the retained set, while invalid or duplicate videos are also removed.

The released WildEvent-1K contains **538,539 videos**, **12 L1 event families**, **129 L2 categories**, and **1,040 leaf event classes**, with one root-to-leaf label per video, as shown in Table 6. The complete set of L1, L2, and leaf event cards is provided separately as supplementary resources.

B Additional Details of Generative Hierarchical Video Classification

This section provides additional implementation details of Generative Hierarchical Video Classification (GHVC) described in Section 4. We further describe the category representation and initialization, autoregressive hierarchical training, hierarchy-constrained inference, and an efficient implementation of answer-only training.

B.1 Category Representation and Token Initialization

As described in Section 4, GHVC represents a root-to-leaf category path either using natural-language category names or dedicated category tokens. For a three-level ground-truth path

$$\pi(V) = (n_1, n_2, n_3), \tag{6}$$

the output can be represented as

$$c(n_1) \rightarrow c(n_2) \rightarrow c(n_3), \tag{7}$$

where $c(n)$ denotes the natural-language name of node n, or as

$$\mathbf{y} = (\tau_{n_1}, \tau_{n_2}, \tau_{n_3}), \tag{8}$$

Table 6: Statistics of WildEvent-1K and representative video classification benchmarks. WildEvent-1K provides a large, explicitly hierarchical event label space with a naturally long-tailed distribution across fine-grained event categories.

(a) Scale and label-space comparison

Dataset	Classification target	Label depth	L1 / L2 / Leaf	Train	Val.	Test	Total
Ours	**Event semantics**	**3**	**12 / 129 / 1,040**	**488,540**	–	**49,999**	**538,539**
K-400 (Kay et al., 2017)	Human actions	1	– / – / 400	246,245	20,000	40,000	306,245
SS V2 (Goyal et al., 2017)	Fine-grained human–object actions	1	– / – / 174	168,913	24,777	27,157	220,847

(b) Distribution across top-level event families

Top-level event family	L2	Leaves	Train	Test	Med. train per leaf	Train share
Spoken Communication	6	63	157,540	6,767	542	32.2%
Object Display	12	90	92,414	7,986	632.5	18.9%
Life and Society	13	251	54,550	9,446	81	11.2%
Making, Repair and Operation	19	132	35,006	5,791	119	7.2%
Vehicle Movement	6	24	31,386	1,807	217	6.4%
Food and Beverage	11	120	28,933	4,688	82	5.9%
Human–Mammal/Bird Interaction	10	48	25,237	2,596	146	5.2%
Sports and Fitness	18	100	22,524	3,527	62	4.6%
Human–Aquatic/Other Interaction	9	40	12,348	1,683	85.5	2.5%
Mammals and Birds	11	52	11,308	1,998	103	2.3%
Scenes and Environments	4	67	11,167	2,465	112	2.3%
Aquatic and Other	10	53	6,127	1,245	23	1.3%
Total	**129**	**1,040**	**488,540**	**49,999**	**105**	**100%**

Notes. Head/medium/tail groups used in the long-tail experiments are *not* obtained by dividing the 1,040 categories into equal thirds. They are defined from the uncapped training frequency of each complete leaf path using the thresholds ≥ 101, 20–100, and ≤ 19, respectively. Under the classification-formulation evaluation set, these thresholds yield 525 head, 357 medium, and 158 tail classes with ground-truth test examples. A dash indicates that WildEvent-1K does not use a separate validation split. Kinetics-400 statistics correspond to the original release.

where τ_n is a dedicated category token associated with taxonomy node n.

We use the category-token representation as the default implementation. Each non-root node has its own token, yielding $1{,}181$ category tokens for WildEvent-1K. This provides a fixed three-token prediction target while preserving a distinct prediction unit for every node in the taxonomy. Tokens are associated with taxonomy nodes rather than category-name strings, so nodes with identical or similar local names under different parents remain distinct.

Because newly introduced category tokens do not have pretrained representations, we initialize each token from the natural-language name of its corresponding category. Let the category name of node n be tokenized into pretrained vocabulary tokens $(w_1, \ldots, w_K)$. Given the input embedding matrix E, we initialize

$$E^{(0)}_{\tau_n} = \frac{1}{K} \sum_{k=1}^{K} E_{w_k}. \tag{9}$$

When the input embedding and language-model output head are not tied, the corresponding output row is initialized analogously:

$$W^{(0)}_{\tau_n} = \frac{1}{K} \sum_{k=1}^{K} W_{w_k}. \tag{10}$$

This initialization provides each newly introduced category token with a semantic starting point derived from the pretrained language space, rather than initializing it as an arbitrary class symbol.

B.2 Autoregressive Hierarchical Training

For each video V, the model receives sampled video frames, a classification instruction I, and the taxonomy $\mathcal{T}$ containing category names and their parent–child relations. Using the default category-token representation, the target for a ground-truth path $\pi(V) = (n_1, n_2, n_3)$ is

$$\mathbf{y} = (\tau_{n_1}, \tau_{n_2}, \tau_{n_3}). \tag{11}$$

GHVC is trained autoregressively with teacher forcing:

$$\mathcal{L}_{\mathrm{GHVC}} = -\sum_{\ell=1}^{3} \log p_\theta \left(\tau_{n_\ell} \mid V, \mathcal{T}, I, \tau_{n_{<\ell}}\right). \tag{12}$$

The Level-2 category is therefore predicted conditioned on the ground-truth Level-1 token, while the leaf category is conditioned on both ground-truth ancestors. The category-name representation follows the same coarse-to-fine procedure using natural-language category names instead of dedicated tokens.

Only the assistant answer positions contribute to the training loss. The sampled video frames, classification instruction, and taxonomy serve entirely as conditioning context and are excluded from token-level supervision.

Importantly, training retains the standard language-model softmax over the extended vocabulary. We do not introduce a separate classification head, sibling-only normalization during training, or an auxiliary hierarchical loss. Instead, hierarchy is incorporated through three components: the taxonomy provided in the context, the coarse-to-fine output sequence, and autoregressive conditioning on ancestor categories.

B.3 Hierarchy-Constrained Inference

At inference time, GHVC follows the same coarse-to-fine autoregressive structure as training, but replaces ground-truth ancestors with previously predicted categories. To ensure that the generated sequence always corresponds to a valid taxonomy path, each prediction is restricted to the children of the predicted parent.

Let $\mathrm{Ch}(p)$ denote the valid children of parent node p, and let r denote the taxonomy root. Starting from $\hat{n}_0 = r$, the prediction at level ℓ is

$$\hat{n}_\ell = \arg \max_{n \in \mathrm{Ch}(\hat{n}_{\ell-1})} z_{\tau_n}^{(\ell)}, \tag{13}$$

where $z_{\tau_n}^{(\ell)}$ is the language-model logit associated with category token τ_n at level ℓ.

After each decision, the selected category token is appended to the prediction prefix before the next level is evaluated. For WildEvent-1K, inference therefore proceeds through three successive predictions:

$$\hat{n}_1 \rightarrow \hat{n}_2 \rightarrow \hat{n}_3. \tag{14}$$

The first prediction selects a Level-1 category. The second prediction considers only its valid Level-2 children, and the final prediction considers only the leaf categories under the predicted Level-2 node.

This procedure decomposes the global classification problem into a sequence of local decisions and guarantees that every output forms a valid root-to-leaf path. Categories outside the selected branch cannot be predicted at deeper levels, eliminating the need for post-hoc hierarchy correction or parsing of generated category names. The same parent–child constraint is applied when using natural-language category names as the output representation.

B.4 Efficient Answer-Only Training

Including the complete taxonomy provides useful semantic and structural context but also makes the conditioning sequence substantially longer than the classification target. Although answer-only loss

masking prevents prompt positions from contributing to the objective, a naive implementation may still project the hidden states of all sequence positions through the full language-model output head.

Let the complete multimodal sequence contain L hidden states, with hidden dimension h, vocabulary size $|\mathcal{V}|$, and batch size B. Applying the output projection to every position has computational complexity proportional to

$$O(BLh|\mathcal{V}|). \tag{15}$$

However, only the hidden states corresponding to the K supervised answer positions are required to compute the causal language-model loss. We therefore apply the output projection only to these positions, reducing the projection cost to

$$O(BKh|\mathcal{V}|), \tag{16}$$

where $K \ll L$. For the default category-token representation of WildEvent-1K, the classification target contains only three category tokens.

This optimization affects only the final vocabulary projection. The complete multimodal sequence—including all sampled video frames, the instruction, the taxonomy, and the autoregressive answer prefix—is still processed by the VLM backbone. Consequently, every supervised prediction retains access to the full conditioning context.

The optimization therefore preserves the same answer-only GHVC objective while avoiding unnecessary full-vocabulary projections at unsupervised positions, reducing both memory usage and computation when a large taxonomy is included in the prompt.

C Additional Experimental Details

This section provides additional details of the datasets, preprocessing, baselines, optimization, and evaluation protocols used in Section 5.

C.1 Datasets and Splits

Our primary experiments are conducted on WildEvent-1K, constructed from VidGen-1M using the EventTree pipeline described in Section 3. After video-quality filtering, taxonomy construction, category verification, and final label assignment, the dataset contains 538,539 videos organized into 12 top-level event families, 129 intermediate categories, and 1,040 leaf event classes. Each video is associated with a unique root-to-leaf path.

To prevent source leakage, clips are grouped by their original source videos, and all clips from the same source are assigned to the same split. The final release contains 488,540 training videos and 49,999 test videos; we do not use a separate validation split. Near-duplicate and corrupted videos are removed before the final split.

We additionally evaluate models on Kinetics-400 (K400) (Kay et al., 2017), which provides a complementary conventional action-recognition setting. Because K400 and WildEvent-1K differ substantially in both category definition and label-space size, we use K400 primarily to compare the relative behavior of model families rather than directly comparing absolute performance across datasets.

C.2 Video Encoder Baselines

We evaluate representative video encoders, including Video Swin Transformer (Liu et al., 2022), VideoMAE (Tong et al., 2022), and InternVideo2 (Wang et al., 2024). For each encoder, the pooled video representation is followed by a linear classification head.

On WildEvent-1K, we evaluate both frozen-backbone linear probing and downstream adaptation. For the controlled encoder–VLM comparison, models use matched downstream data and a single temporal clip and spatial view. We initialize video encoders from publicly released pretrained checkpoints and follow their recommended downstream optimization recipes. Because the compared encoders differ in architecture, parameter count, pretraining objectives, and pretraining data, this experiment should be interpreted as a comparison between pretrained model families under the same downstream task rather than as a controlled study of individual pretraining factors.

C.3 VLM Classification Formulations

We use Qwen3.5 models (Team, 2026) as the primary VLM family and compare four formulations that separate the effects of visual representation, category semantics, and hierarchical prediction.

VLM + Linear Head. We pool the hidden states corresponding to video tokens and apply a conventional linear classifier over the legal leaf categories. The predicted leaf is then mapped to its unique three-level taxonomy path. This baseline does not use language-based category semantics during prediction.

Flat Name Generation. The VLM generates the three category names as free text using its original language vocabulary, without taxonomy snapping or parent-conditioned decoding. The generated fields are parsed into an L1–L2–leaf path. This formulation exposes natural-language category semantics but does not structurally enforce a valid taxonomy path.

Flat Category Token. Each leaf category is represented by a dedicated semantic category token initialized from its natural-language label. Prediction is restricted to the legal leaf-token set, after which the selected leaf is mapped to its unique three-level path. This removes variable-length label generation while retaining a flat global decision over the leaf label space.

GHVC. Our method represents taxonomy nodes using semantic category tokens and predicts the complete category path coarse-to-fine. During inference, each finer-level prediction is restricted to the valid children of the selected parent, as described in Section 4.

For zero-shot evaluation, we use natural-language category names from the pretrained vocabulary rather than newly introduced category tokens.

C.4 Video Preprocessing

For the primary VLM experiments, frames are sampled at 2 fps, with the number of frames clamped between 8 and 24. True frame timestamps are retained and provided to the Qwen3.5 processor. The default total spatial budget is 2,097,152 pixels per video, corresponding to approximately 960 video tokens. Audio is not used.

For the controlled video-encoder comparison in Section 5.2, video encoders use the frame settings reported in Table 1. Evaluation uses a single temporal clip and a single spatial view unless otherwise specified.

C.5 VLM Training Details

VLMs are trained for two epochs. The per-device batch size is 4 with two gradient-accumulation steps across eight GPUs, giving a global batch size of 64. We use AdamW with cosine learning-rate decay, a warm-up ratio of 0.03, and bfloat16 training.

Our default parameter-efficient configuration applies LoRA to the q, k, v, and output projections of the language-model attention layers while keeping the vision tower frozen. The embeddings and output rows corresponding to newly introduced semantic category tokens remain trainable. We use a LoRA rank of $r = 16$, a LoRA scaling factor of $\alpha = 32$, and a learning rate of 1×10^{-4}.

Full-model fine-tuning uses a learning rate of 1×10^{-5}. LoRA experiments use DeepSpeed ZeRO-2, whereas full-model fine-tuning uses ZeRO-3. For language-model-based classification, loss is applied only to supervised answer tokens. Video tokens, taxonomy text, and task instructions are used as conditioning context and are masked from the language-modeling loss. The efficient answer-only output projection used by GHVC is described in Appendix B.

C.6 Evaluation Metrics

Accuracy and macro-F1. For K400, we report standard Top-1 accuracy. For WildEvent-1K, each leaf class is identified by its complete L1–L2–leaf path. We report Top-1 accuracy at all three taxonomy levels, denoted by Acc_{L1}, Acc_{L2}, and Acc_{Leaf}. We additionally report leaf macro-F1.

Per-class precision, recall, and F1 are computed from exact full-path predictions over the complete evaluation set, and macro-F1 gives equal weight to every leaf category represented by ground-truth examples in that evaluation set. Invalid or incomplete predictions remain in the evaluation records and therefore contribute false negatives to their ground-truth classes.

Tree distance. TreeDist measures the semantic severity of an error by the shortest-path distance between the predicted and ground-truth leaves in the fixed three-level taxonomy. Let $\mathcal{C}$ be the set of legal complete leaf paths and let $d_{\mathcal{T}}$ denote the number of tree edges between two legal leaves. For sample i, we define

$$D_i = \begin{cases} d_{\mathcal{T}}(\hat{y}_i, y_i), & \hat{y}_i \in \mathcal{C}, \\ 6, & \hat{y}_i \notin \mathcal{C}, \end{cases} \qquad \text{TreeDist} = \frac{1}{N}\sum_{i=1}^{N} D_i. \tag{17}$$

Because all leaves are at depth three, legal predictions have distances in $\{0, 2, 4, 6\}$ according to their deepest shared ancestor with the ground-truth leaf. Invalid predictions—including missing fields, unknown labels, or illegal parent–child combinations—receive the maximum penalty of 6. TreeDist is therefore an unnormalized average in $[0, 6]$; lower is better.

Hierarchy consistency and parse failure. Hierarchy consistency is the fraction of evaluated predictions that form a legal complete path in the taxonomy:

$$\text{Consistency} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}[\hat{y}_i \in \mathcal{C}]. \tag{18}$$

Parse failure is the fraction of predictions for which at least one of the three output fields is missing or empty:

$$\text{ParseFail} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}[\text{at least one predicted field is missing}]. \tag{19}$$

The two quantities are distinct: a generated output can contain three successfully parsed labels yet still form an invalid taxonomy path. Such a case has zero parse failure, zero hierarchy consistency for that sample, and TreeDist $= 6$. In the classification-formulation study, the linear head, flat category token, and hierarchical token formulations produce legal paths by construction when prediction succeeds, whereas Flat Name Generation is parsed from unconstrained three-level text and uses no taxonomy snapping.

Head/medium/tail F1. Long-tail groups are defined from the number of examples in the original, uncapped training annotations for each complete leaf path. We use fixed thresholds rather than equal-sized groups:

$$\mathcal{C}_H = \{c : n_c^{\text{train}} \geq 101\}, \qquad \mathcal{C}_M = \{c : 20 \leq n_c^{\text{train}} \leq 100\}, \qquad \mathcal{C}_T = \{c : n_c^{\text{train}} \leq 19\}. \tag{20}$$

We first compute TP_c, FP_c, FN_c, and $F1_c$ for every legal leaf category on the *full evaluation set*. We then macro-average $F1_c$ within the head, medium, and tail groups, considering only categories with at least one ground-truth example in the corresponding evaluation set. This order preserves cross-group false positives; we do not filter the evaluation videos by group before computing per-class F1. Under the current classification-formulation evaluation set, the three groups contain 525, 357, and 158 evaluated classes, respectively.

C.7 Evaluation Protocol

All methods within a reported comparison are evaluated on the same available-video set. Distributed predictions are gathered before the metrics are computed; scores are not obtained by averaging per-GPU metrics. Videos unavailable because of missing frame caches are removed when constructing the common evaluation set and are not counted as parse failures.

For the classification-formulation experiment in Table 2, all four methods are evaluated on the same 49,895 videos: 49,999 videos belong to the released test split, and 104 videos with missing frame caches are excluded before evaluation. Flat Name Generation is evaluated without taxonomy

snapping. The structured formulations use their corresponding constrained output spaces described above. Diagnostic options that limit the number of evaluated samples or the number of examples per class are not used for the reported results.

For model-scaling and long-tail comparisons, the same metric definitions are applied after first determining the shared available-video set for the models being compared. Consequently, the reported N is a property of the corresponding evaluation suite rather than a fixed part of the metric definitions.