EthoCLIP: Ontology-Enhanced Video-Language Pretraining for Animal Behavior Understanding

Supplementary Material

1. Additional Related Work

1.1. Animal Behavior Ontology

An animal behavior ontology is a formal, machine-readable ontology that provides a controlled vocabulary and logical structure for describing animal behaviors and their relations, with the goal of enabling consistent annotation, integration, and computational analysis of behavioral data across studies and taxa [1–3]. It aims to support consistent annotation and cross-study comparison. In practice, existing resources fall broadly into species-specific ethograms and multi-species ontologies or lexica, each with distinct limitations. For specific-species ethogram, Felidae ethogram propose "base behaviors" plus modifiers to harmonize terminology across felid studies while preserving species-level detail [1]. Likewise, the African Elephant Ethogram developed by Poole and colleagues catalogues more than 300 behaviors and over 100 behavioral suites, with textual definitions and over 2,400 audio–visual exemplars for savanna elephants [2]. These resources are ethologist-driven and behaviorally fine-grained, but they are inherently taxon-bound: label sets, categories and granularity are tailored to one clade and do not align systematically across species, making cross-species transfer and large-scale integration difficult. Otherwise, multi-species lexical resources, such as "A Dictionary of Animal Behavior" [3] offer a concise, cross-taxa vocabulary. The dictionary contains over 480 entries and explicitly incorporates terms from ecology, physiology, and psychology, making it a broad conceptual reference rather than a dataset-oriented label set. However, despite this breadth, it remains essentially a textual lexicon rather than a formal ontology: entries are organized alphabetically, but without an explicit hierarchy of relations, or any direct alignment to contemporary video annotations. Moreover, a non-trivial subset of entries refers either to behavioral phenomena that are primarily elicited in controlled laboratory settings, or to internal physiological and cognitive processes that cannot be inferred reliably from external observation alone. Hence, mapping real-world, visually grounded labels from camera-trap or field video datasets to such dictionary terms is often ambiguous and subject to systematic mismatches in observability and granularity.

1.2. Vision-Language Learning

In recent years, vision-language pretraining models (VLMs) have demonstrated strong zero-shot generalization across a variety of downstream tasks, becoming a key trend in computer vision and multimodal learning [4–10]. Representative models include ALIGN [9], CLIP [8], and BLIP [10], which are typically trained on large-scale datasets containing billions of image-text pairs. During training, VLMs employ a contrastive learning objective to jointly optimize the image encoder and text encoder, ensuring effective alignment between visual and textual modalities in a shared embedding space [11, 12]. During inference, these models embed category names or textual descriptions into a linear classifier to enable zero-shot predictions. Despite their success in image recognition, video understanding, and multimodal retrieval tasks [13–15], VLMs have shown remarkable generalization without task-specific fine-tuning. Meanwhile, VLMs have also achieved significant progress across various specialized domains. In the medical domain, medical large models such as ChiMed-GPT [16], AlpaCare [17], and Taiyi [18] have been applied to medical question answering, diagnostic assistance, and professional knowledge retrieval. In the financial domain, models such as DISC-FinLLM [19] support financial analysis and intelligent question answering. In the education domain, models like Taoli [20] and EduChat [21] facilitate intelligent tutoring and personalized learning. These results indicate that large models trained with contrastive learning or multimodal pretraining exhibit strong generalization and adaptability across domains. However, their application in animal behavior recognition remains largely unexplored, highlighting a promising direction for future research.

1.3. Graph Modeling

Recent research has explored graph-based modeling to encode structured behavior and label spaces, facilitating hierarchical representation and consistent annotation across datasets [22–25]. HGCLIP integrates a class taxonomy into a graph and applies Graph Neural Networks (GNNs) to propagate semantic information along parent–child relations [24]. This approach captures coarse-to-fine hierarchical dependencies, enhancing generalization across granularity levels and improving classification performance in fine-grained and hierarchical settings. To address cross-dataset label inconsistencies, O'Neil et al. construct a unified ontology that maps semantically related labels through hierarchical graph structures [23]. Such a shared semantic backbone enables consistent annotation, reduces ambiguity, and improves classification via semantic regularization. ChimpVLM leverages a standardized ethogram and

language-model embeddings to encode textual behavior descriptions, initializing transformer class tokens according to semantic proximity [22]. This strategy enhances recognition, particularly for rare behaviors. Finally, Gao et al. employ a knowledge graph linking action labels, attributes, and object concepts, with edges encoding relations such as "has-attribute" or "similar-to" [25]. Parallel GCN streams jointly reason over visual inputs and semantic structure, enabling zero-shot inference for unseen actions. Collectively, these studies demonstrate that graph-based modeling—through explicit ontologies, hierarchical taxonomies, or language-derived semantic embeddings—effectively captures relational and semantic structures, facilitating label unification, robust generalization, and performance improvements in low-resource or zero-shot settings.

2. Details of AnimalBand

2.1. Semantic Distribution

Fig. 1 presents the number of samples for standardized behaviors in the AnimalBand dataset, along with detailed statistics for each category. Common daily behaviors, such as *chewing* and *resting posture*, are highly represented, whereas *pleasure behavior* is considerably less frequent. This distribution difference mainly arises from variations in occurrence frequency and observability across behaviors. For instance, *chewing* is a high-frequency and sustained daily behavior, making it easier to be captured in videos and accurately annotated, whereas *pleasure behavior* occurs less frequently in natural settings and has less visually salient features, making it more likely to be overlooked during annotation. This distribution pattern broadly reflects the natural occurrence of animal behaviors, where routine and repetitive actions dominate, while rare or brief behaviors are comparatively infrequent.

2.2. Details of Data Mapping

In this study, we employ the GPT-5 API to perform automated label mapping and standardization across multiple datasets. During the formation of AnimalBand, behavioral labels from different experiments and publicly available datasets are often heterogeneous, exhibiting differences in naming conventions, granularity, and semantic ambiguity. To unify these labels, we leverage GPT-5's powerful semantic understanding, natural language reasoning, and cross-modal alignment capabilities to map heterogeneous labels onto a unified, manually curated Neurobehavioral Ontology (NBO), as illustrated in Table 1.

The procedure is as follows. First, GPT-5's operational principles are globally set through the API to ensure consistent and controlled behavior throughout the label mapping process. The label mapping is performed according to several core principles: semantic similarity is prioritized to maximize alignment between candidate labels and the target label's meaning, using available definitions and synonyms; lexical or lemma overlap is considered when semantics align, favoring candidates sharing key words with the verb; granularity and prototypicality are taken into account to select moderately specific and representative labels while avoiding overly broad or overly fine-grained options; common usage is considered to prefer labels that are most frequently used in the target domain; observability is used as a tie-breaker, with directly observable behaviors preferred over abstract states when semantics are comparable; and in cases where multiple candidates still satisfy the above criteria, the most semantically prototypical option is chosen. Example input-output pairs are provided to clarify the expected mapping procedure. For instance, the verb *hunts other animals* is mapped to the label *predator behavior*, generating the JSON output {*"match": "predator behavior"*}.

Finally, all automated mappings are manually verified and corrected to ensure accuracy and consistency. Through this process, we obtain a reliable and standardized semantic representation, enabling cross-dataset animal behavior recognition. This approach improves label mapping efficiency, providing a solid foundation for the construction of AnimalBand.

2.3. Data License/Address

All data in AnimalBand are obtained from publicly available datasets and are intended solely for research purposes. The licenses and links to the original datasets are provided below:

- **Animal Kingdom**
  Link: https://github.com/sutdcv/Animal-Kingdom
  License: https://docs.google.com/forms/d/e/1FAIpQLSccqs92uqbaf1jfHrWB7lMejCDp8galelSRhp4u9DpX5-x3YQ/viewform
- **MammalNet**
  Link: https://mammal-net.github.io/
  License: https://creativecommons.org/licenses/by/4.0/
- **LoTE-Animal**
  Link: https://lote-animal.github.io/
  License: https://creativecommons.org/licenses/by-sa/4.0/
- **PanAf20K**
  Link: https://data.bris.ac.uk/data/dataset/1h73erszj3ckn2qjwm4sqmr2wt
  License: https://www.nationalarchives.gov.uk/doc/non-commercial-government-licence/version/2/
- **MammalAps**
  Link: https://zenodo.org/records/1504090

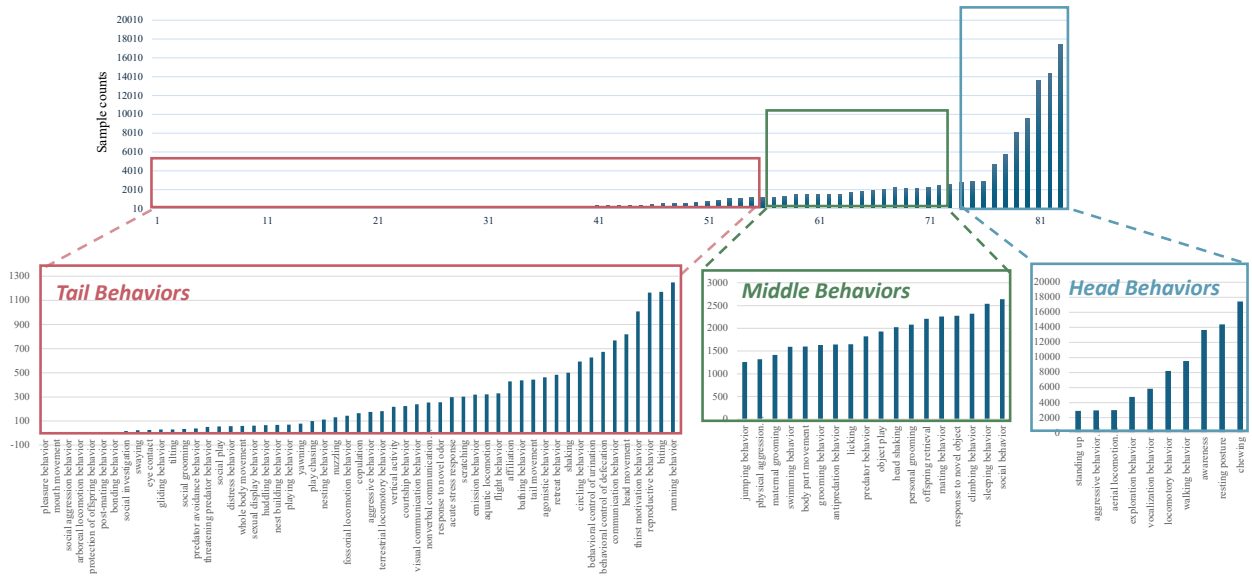


Figure 1. AnimalBand label distribution

1

License: https://creativecommons.org/licenses/by-nc/4.0/legalcode

3. Details of Experiments

3.1. Datasets

ChimpACT: ChimpACT is a video dataset centered on the natural social behaviors of chimpanzees, distinguished by its long-term continuous recordings of a real social group. It spans interactions among more than 20 individuals from 2015 to 2018, capturing rich variations in social dynamics and complex inter-individual relations, which pose substantial challenges for modeling group behavior and long-duration activity patterns. The original annotations of this dataset are designed for video tracking tasks. In our study, we converted it into a multi-label, video-level annotation format following the style of Kinetics-400. All videos are randomly divided into training, validation, and test sets using a 7:1:2 ratio.

SheepActivity: The SheepActivity dataset is characterized by its multi-view and multi-resolution recordings of both flock-level and individual sheep activities, making it well suited for investigating cross-view behavior recognition. The dataset contains a balanced set of canonical grazing-related behaviors and supports evaluating model robustness under diverse camera conditions. We split the dataset into training, validation, and test sets with a 7:1:2 ratio.

Panda: The Wolong wild giant panda dataset is notable for its authentic wildlife footage captured in natural habitats, featuring substantial environmental complexity such as illumination variation and vegetation occlusion. These conditions make behavior recognition significantly more challenging and realistic for ecological monitoring applications. The videos depict typical daily behaviors of wild giant pandas and exhibit clear temporal patterns. In our experiments, the dataset is randomly partitioned into training, validation, and test sets following a 7:1:2 ratio.

3.2. Training Details of EthoCLIP

EthoCLIP uses a frozen CLIP backbone ViT-B/16 to extract visual-language embeddings with high semantic quality. During training, only the spatiotemporal modeling module and the ontology-aware graph modeling component are updated while the parameters of CLIP remain fixed to retain the pretrained alignment between frames and text. The model is trained for 30 epochs using the AdamW optimizer with a cosine learning rate schedule and linear warmup for the first five epochs. The base learning rate is 8e-6 and the minimal learning rate is 8e-8. The batch size per GPU is 16 with four accumulation steps. Weight decay is set to 0.001. Data augmentation includes label smoothing, Mixup and CutMix. A detailed summary of all training parameters is provided in Table 2.

3.3. Training Details of Traditional Models

We conduct experiments using the MMAction2 [26] and PySlowFast [27] codebases, following their recommended training protocols and hyperparameter settings. Each model is trained for 50 epochs. For ActionCLIP and UniFormer v2, we initialize the models with weights pretrained on JFT and Kinetics-710. During training, each video is uniformly sampled into clips of 8 consecutive frames. During validation and testing, we adopt a multi-view evaluation protocol

Table 1. Instructions for using a large language model to perform label mapping when forming AnimalBand.

A. Specify the core matching principles:

You are an NLP behavior matcher. Given a single verb and a list of candidate behaviors, choose EXACTLY ONE item that best matches the verb's meaning. Output constraints: Return ONLY a JSON object with the key *match*. The value must be copied verbatim from the provided candidate list (case & spacing preserved). Pick exactly one, never abstain, never invent, never modify tokens, never add any other keys or text.

Decision rubric (priority order):

1. Semantic similarity: Maximize closeness between the verb's meaning and each candidate's meaning. If available, use provided verb definition / synonyms and candidate definitions to ground meaning.

2. Lexical/lemma overlap: Prefer candidates sharing lemma/morphology or key content words with the verb when semantics align.

3. Granularity match: Prefer candidates whose specificity matches the verb (neither too broad nor too narrow).

4. Common usage: Prefer the candidate that reflects the most commonly understood usage in the target domain/corpus.

5. Observability: When semantics tie, prefer observable behaviors over states (unless the verb denotes a state).

6. Tie-breaking: If still tied, choose the most semantically prototypical option.

If none is perfect, still choose the closest per the rubric above.

B. Provide the input-output format:

Choose the best matching behavior for the verb below.

Verb: *verb*

Candidate behaviors (comma-separated): *ontology behaviors*

(Optional auxiliary info: provide if available, otherwise omit, verb definition: *def*, verb synonyms: *syns*)

Return only JSON: {"match": "<one item from the list above, copied verbatim>"}

Do not add any explanation.

C. Provide an example for reference:

Example:

Choose the best matching behavior for the verb below.

Verb: hunts other animals

Candidate behaviors (comma-separated): *biting, tongue movement, head rotation, predator behavior, running behavior, climbing behavior, gliding behavior, blinking, vertical activity, sitting down*

Return only JSON: {"match": "<one item from the list above, copied verbatim>"}

Do not add any explanation.

Example output:

{"match": "predator behavior"}

with 4 temporal clips and 3 spatial crops per clip. All models are optimized using AdamW with a cosine learning-rate schedule and an early-stage warm-up.

3.4. VLM Evaluation

We evaluate VideoLLaMA3, InternVL3.5, and Qwen2.5-VL on the test sets of three downstream datasets using their official implementations. To ensure a fair comparison, all models are configured with the same inference settings as shown in Table 3. To accommodate the input requirements of these models, we design dataset-specific question–answer formats. Specifically, we first use GPT to generate ten question templates and randomly select one during inference. For single-label classification tasks, we construct the dataset's label set as the collection of answer options and prompt the model to choose the most appropriate one. For multi-label tasks, each category is treated as an independent query. For every label, we design a question regarding the intensity of the corresponding behavior, and the model outputs a probability ranging from 0 to 1. mAP is adopted as the evaluation metric.

4. More Results

4.1. Results on Mouse Dataset

We also conducted experiments on the CalMS21 dataset of laboratory mice. The original annotations of this dataset are designed for video tracking tasks. In our study, we

Table 2. Full-supervision training configuration for AnimalBand train set.

Component	Setting
Number of frames	8
Number of classes	160
CLIP Model	ViT-B/16
Batch size per GPU	16
Accumulation steps	4
Epochs	30
Optimizer	AdamW
Optimizer betas	(0.9, 0.98)
Learning rate schedule	cosine
Linear warmup epochs	5
Base learning rate	8e-6
Minimal learning rate	8e-8
Weight decay	0.001
Label smoothing	0.1
Mixup	0.8
Cutmix	1.0

Table 3. VLM Evaluation Settings

Parameters	Setting
Max tokens	512
Prompt	Carefully watch the videos/images and pay attention to the cause and sequence of events, the detail and movement of animal, and the action and pose of animal. Based on your observations, select the best option that accurately addresses the question. Please think step by step. Only give the best option. `Best option:(`
Temperature	0.0
Video FPS	1.0
Max pixels	224×224
Max frames	8

converted it into a multi-label, video-level annotation format following the style of Kinetics-400. As shown in 4, under the same settings, our method consistently outperforms existing approaches, achieving high accuracy. These results demonstrate the potential of our method for laboratory mouse behavior analysis, highlighting its applicability in AI-driven scientific studies.

4.2. Ablation Results on Ontology Granularity

We conducted an ablation study on the granularity levels of the ontology. As Table 5, increasing the granularity from 2 to 6 (the maximum) consistently improves accuracy across all categories, suggesting that finer ontology hierarchies provide richer semantic guidance.

Table 4. Comparison on the CalMS21 dataset.

Models	Acc@1
SlowFast	83.16
X3D	90.55
ActionCLIP	92.00
UniFormer v2	95.28
EthoCLIP	**96.50**

Table 5. Ablation study on ontology granularity.

Layers	Head	Middle	Tail	Overall
2	82.50	70.63	50.45	61.30
4	82.72	70.54	50.51	61.36
6	82.54	71.12	53.27	**62.98**

4.3. More Prediction Results

In Figure 2 and 3, we present the prediction results of our method and the baselines, further validating the effectiveness of our method.

EthoCLIP-AnimalBand: **Resting**
Baseline-AnimalBand: **Walking**
Baseline-Merge: **ScentMarking**

EthoCLIP-AnimalBand: **Walking**
Baseline-AnimalBand: **Sniffing**
Baseline-Merge: **ScentMarking**

EthoCLIP-AnimalBand: **ScentMarking**
Baseline-AnimalBand: **ScentMarking**
Baseline-Merge: **Resting**

EthoCLIP-AnimalBand: **Walking**
Baseline-AnimalBand: **Walking**
Baseline-Merge: **Resting**

EthoCLIP-AnimalBand: **Sniffing**
Baseline-AnimalBand: **Sniffing**
Baseline-Merge: **Resting**

EthoCLIP-AnimalBand: **Resting**
Baseline-AnimalBand: **Resting**
Baseline-Merge: **ScentMarking**

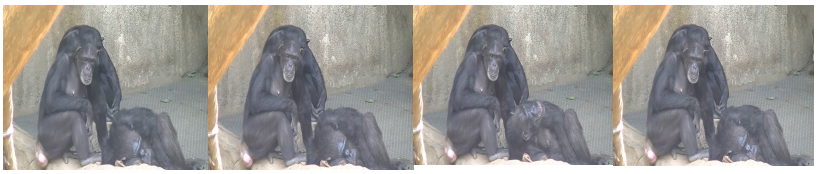
EthoCLIP-AnimalBand: **Walking**
Baseline-AnimalBand: **Walking**
Baseline-Merge: **Sniffing**

Figure 2. Visualization results on the Panda dataset

EthoCLIP-AnimalBand: **Running**
Baseline-AnimalBand: **Walking**
Baseline-Merge: **Sitting**

EthoCLIP-AnimalBand: **Standing**
Baseline-AnimalBand: **Sitting**
Baseline-Merge: **Sitting**

EthoCLIP-AnimalBand: **Running**
Baseline-AnimalBand: **Running**
Baseline-Merge: **Walking**



GT: resting
EthoCLIP-AnimalBand: **resting**, climbing, sleeping, manipulating object, playing
Baseline-AnimalBand: sleeping, climbing, eating, nursing, playing
Baseline-Merge: climbing, losing object, being nursed, moving, eating

GT: climbing, resting, sleeping, manipulating object, grooming, being groomed
EthoCLIP-AnimalBand: **climbing, sleeping, resting**, playing, **manipulating object**
Baseline-AnimalBand: **sleeping, climbing**, playing, **grooming**, eating
Baseline-Merge: **climbing**, losing object, moving, being nursed, eating

Figure 3. Visualization results on the SheepActivity and ChimpACT dataset

References

[1] Lauren Ashley Stanton, Matthew Stephen Sullivan, and Jillian Marie Fazio. A standardized ethogram for the felidae: A tool for behavioral researchers. *Applied Animal Behaviour Science*, 173:3–16, 2015. ISSN 0168-1591. SI: Cats have many lives. 1

[2] Joyce H. Poole and Petter Granli. The elephant ethogram: a library of african elephant behaviour. *Pachyderm*, 62:105–111, 2021. doi: 10.69649/pachyderm.v62i.462. 1

[3] David McFarland. *A Dictionary of Animal Behaviour*. Oxford Quick Reference. Oxford University Press, Oxford, 2 edition, 2014. ISBN 9780191761577. 1

[4] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *Proceedings of the IEEE Int'l Conf. on Computer Vision*, pages 2630–2640, 2019. 1

[5] Chen Sun, Fabien Baradel, Kevin Murphy, and Cordelia Schmid. Learning video representations using contrastive bidirectional transformer. *arXiv preprint arXiv:1906.05743*, 2019.

[6] Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. Videobert: A joint model for video and language representation learning. In *Proceedings of the IEEE Int'l Conf. on Computer Vision*, pages 7464–7473, 2019.

[7] Linchao Zhu and Yi Yang. Actbert: Learning global-local video-text representations. In *Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition*, pages 8746–8755, 2020.

[8] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *Int'l Conf. on Machine Learning*, pages 8748–8763. PMLR, 2021. 1

[9] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *Int'l Conf. on Machine Learning*, pages 4904–4916, 2021. 1

[10] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR, 2022. 1

[11] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PmLR, 2020. 1

[12] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*, 2022. 1

[13] Samuel Stevens, Jiaman Wu, Matthew J Thompson, Elizabeth G Campolongo, Chan Hee Song, David Edward Carlyn, Li Dong, Wasila M Dahdul, Charles Stewart, Tanya Berger-Wolf, et al. Bioclip: A vision foundation model for the tree of life. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19412–19424, 2024. 1

[14] Antoine Miech, Jean-Baptiste Alayrac, Lucas Smaira, Ivan Laptev, Josef Sivic, and Andrew Zisserman. End-to-end learning of visual representations from uncurated instructional videos. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9879–9889, 2020.

[15] Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. Flava: A foundational language and vision alignment model. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15638–15650, 2022. 1

[16] Yuanhe Tian, Ruyi Gan, Yan Song, Jiaxing Zhang, and Yongdong Zhang. Chimed-gpt: A chinese medical large language model with full training regime and better alignment to human preferences. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7156–7173, 2024. 1

[17] Xinlu Zhang, Chenxin Tian, Xianjun Yang, Lichang Chen, Zekun Li, and Linda Ruth Petzold. Alpacare: Instruction-tuned large language models for medical application. *arXiv preprint arXiv:2310.14558*, 2023. 1

[18] Ling Luo, Jinzhong Ning, Yingwen Zhao, Zhijun Wang, Zeyuan Ding, Peng Chen, Weiru Fu, Qinyu Han, Guangtao Xu, Yun-Zhong Qiu, Dinghao Pan, Jiru Li, Hao Li, Wenduo Feng, Senbo Tu, Yuqi Liu, Zhihao Yang, Jian Wang, Yuanyuan Sun, and Hongfei Lin. Taiyi: A bilingual fine-tuned large language model for diverse biomedical tasks. *Journal of the American Medical Informatics Association : JAMIA*, 2023. URL https://api.semanticscholar.org/CorpusID:265294661. 1

[19] Wei Chen, Qiushi Wang, Zefei Long, Xianyin Zhang, Zhongtian Lu, Bingxuan Li, Siyuan Wang, Jiarong Xu, Xiang Bai, Xuanjing Huang, and Zhongyu Wei. Disc-finllm: A chinese financial large language model based on multiple experts fine-tuning, 2023. URL https://arxiv.org/abs/2310.15205. 1

[20] Jingsi Yu, Junhui Zhu, Yujie Wang, Yang Liu, Hongxiang Chang, Jinran Nie, Cunliang Kong, Ruining Chong, XinLiu, Jiyuan An, Luming Lu, Mingwei Fang, and Lin Zhu. Taoli llama. https://github.com/blcuicall/taoli, 2023. 1

[21] Yuhao Dan, Zhikai Lei, Yiyang Gu, Yong Li, Jianghao Yin, Jiaju Lin, Linhao Ye, Zhiyan Tie, Yougen Zhou, Yilei Wang, et al. Educhat: A large-scale language model-based chatbot system for intelligent education. *arXiv preprint arXiv:2308.02773*, 2023. 1

[22] David Brookes, Aileen Li, Panos Achlioptas, Li Fei-Fei, and Jiajun Wu. Chimpvlm: Vision-language modeling for primate behavior understanding with ethogram priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 1, 2

[23] Michael O'Neil, Lei Zhang, and Xin Huang. Unified ontology mapping for cross-dataset semantic consistency. *Journal of Machine Learning Research*, 25(114):1–29, 2024. 1

[24] Cheng Peng, Shuchang Zheng, Zhaoyang Cui, Lin Du, Zhilin Jin, Wei Li, Junjie Yan, and Yao Zhou. Hgclip: Hierarchy-guided contrastive learning for hierarchical image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 1

[25] Chenyou Gao, Yixin Zou, Jia-Bin Huang, and Larry S Davis. Knowledge graphs for zero-shot action recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 8303–8310, 2019. 1, 2

[26] MMAction2 Contributors. Openmmlab's next generation video understanding toolbox and benchmark. https://github.com/open-mmlab/mmaction2, 2020. 3

[27] Haoqi Fan, Yanghao Li, Bo Xiong, Wan-Yen Lo, and Christoph Feichtenhofer. Pyslowfast. https://github.com/facebookresearch/slowfast, 2020. 3