# ShotVL: Human-Centric Highlight Frame Retrieval via Language Queries

**Wangyu Xue**[*1], **Chen Qian**[*1,2], **Jiayi Wu**[2], **Yang Zhou**[2], **Wentao Liu**[2], **Ju Ren**[†1], **Siming Fan**[‡2], **Yaoxue Zhang**[1]

[1]Department of Computer Science and Technology, Tsinghua University
[2]SenseTime Research

## Abstract

Existing research on human-centric video understanding typically focuses on analyzing specific moments or entire videos. However, many applications require higher precision at the frame level. In this work, we propose a novel task, BestShot, which aims to locate highlight frames within human-centric videos through language queries. This task requires not only a deep semantic understanding of human actions but also precise temporal localization. To support this task, we introduce the BestShot Benchmark. The benchmark is meticulously constructed by combining human-annotated highlight frames, duration labels and detailed textual descriptions. These descriptions cover three critical elements: (1) Visual content; (2) Fine-grained actions; and (3) Human pose descriptions. Together, these elements provide the necessary precision to identify the exact highlight frames in videos.

To tackle this problem, we have collected two distinct datasets: (i) ShotGPT4o Dataset, which is algorithmically generated by GPT-4o and (ii) Image-SMPLText Dataset, which features large-scale and accurate per-frame pose descriptions using PoseScript and existing pose estimation datasets. Based on these datasets, we present a strong baseline model, ShotVL, fine-tuned from InternVL, specifically for BestShot. We highlight the impressive zero-shot capabilities of our model and offer comparative analyses with existing state-of-the-art (SOTA) models. ShotVL demonstrates a significant 64% improvement over InternVL on the BestShot Benchmark and a notable 68% improvement on the THUMOS14 Benchmark, while maintaining SOTA performance in general image classification and retrieval.

**Code** — https://github.com/ShotVL/ShotVL

**Extended version** — https://arxiv.org/abs/2412.12675

## Introduction

Within the field of video understanding, locating highlight frames in human-centric videos using language queries is a crucial yet underexplored task. Many downstream tasks heavily rely on frame-level understanding to function effectively. For stepwise video captioning, it is essential to generate accurate video captions that align with specific actions and timestamps, such as in sports ("a player scores a goal between 00:01:05 and 00:01:08") or in instructional videos ("the chef flips the pancake at exactly 00:02:03"). In moment retrieval, precise frame localization ensures that the retrieved moments exactly match the query, particularly for short-duration and complex actions.

| Dataset | Domain | #Query | Duration | #Action | #Video | Avg word |
|---|---|---|---|---|---|---|
| AVA-2.2 | Movie | 80 | 1s | - | 430 | 2 |
| THUMOS14 | Sports | 20 | 4.3s | 20 | 413 | 3 |
| FAction | Daily | 106 | 7.1s | 106 | 16.7K | 2 |
| MSports | Sports | 66 | 1.0s | 4 | 0.8K | 2 |
| HiRest | Daily | 8K | 7.6s | - | 3.4K | 4 |
| DiDeMo | Flickr | 41K | 8.0s | - | 10.6K | 7 |
| ANet | Activity | 72K | 7.2s | - | 15K | 14 |
| Charades | Activity | 16K | 13.4s | - | 6.7K | 6 |
| QVHL | Vlog/ News | 10K | 11.3s | - | 10.2K | 11 |
| **BestShot** | TH14, FAction | 6K | **12 frames** | 59 | 628 | 78 |

Table 1: Benchmark and Dataset Comparison: TH14 (THUMOS14), FAction (FineAction), MSports (MultiSports). BestShot is only for zero-shot test.

Despite the importance of precise frame-level localization, existing methods face significant limitations, mainly due to insufficient granularity and inadequate language descriptions. Many current approaches focus on broad temporal segments instead of precise frames, which limits their effectiveness in tasks that require detailed understanding. Furthermore, the restricted length of their language descriptions, usually ten words or fewer, hinders the retrieval of frames with complex or nuanced descriptions. These methods also struggle to generalize across diverse scenes and actions, reducing their reliability in real-world applications, where content is typically more varied and unpredictable.

In this paper, we introduce a challenging benchmark, the BestShot Benchmark. This benchmark aims to address the limitations outlined above by offering a more granular evaluation metric that more accurately reflects the demands of high-precision frame understanding. The benchmark facilitates more convenient and reliable evaluation of mod-

[*]These authors contributed equally.
[†]Corresponding Author, renju@tsinghua.edu.cn
[‡]Corresponding Author, gzfansiming@gmail.com

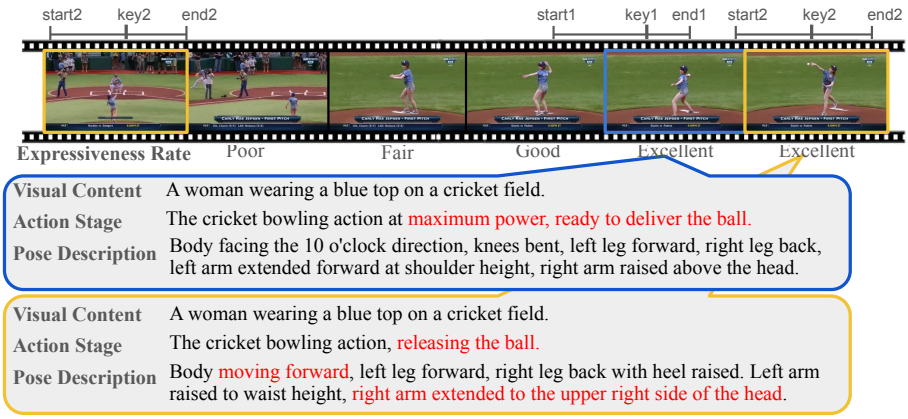


Figure 1: Example of BestShot Benchmark. The task is to locate the exact frame through language queries related to content, action stage and pose description. Each query may correspond to multiple intervals.

els' ability to precisely localize highlight frames in human-centric videos using detailed queries.

The BestShot Benchmark reveals several key challenges that must be addressed: (1) Data Limitations: Existing large-scale image and video captioning datasets often emphasize coarse-grained summaries rather than detailed descriptions. Fine-grained annotations, such as ShareGPT4V (Chen et al. 2023a), are limited in both scale and precision, particularly when it comes to capturing detailed action stages and pose descriptions. Furthermore, most moment-retrieval datasets are manually annotated and relatively small, which restricts the generalization capabilities of models trained on them. (2) Method Limitations: The performance of current baseline methods on the proposed benchmark remains sub-par. Large vision-based models, such as InternVL (Chen et al. 2023b) and InternVideo (Wang et al. 2022), struggle to localize frames in videos without fine-tuning domain-specific data. Additionally, while video LLMs such as LITA (Huang et al. 2024) and VTG-LLM (Guo et al. 2024) exhibit some localization and temporal understanding capabilities, their generalization is hindered by insufficient training data.

Therefore, in alignment with the proposed benchmark, we introduce two large-scale training datasets relevant to BestShot, along with a robust baseline model, ShotVL, to tackle these challenges. Our primary contributions are summarized as follows:

- We introduce the BestShot Benchmark for precise highlight frame retrieval in human-centric videos using complex and fine-grained language queries.
- We propose the large-scale ShotGPT4o Dataset and Image-SMPLText Dataset to address the challenges of limited data diversity and insufficient fine-grained descriptions.
- We provide a comprehensive evaluation of existing methods and propose a robust baseline ShotVL. ShotVL demonstrates a significant 64% improvement over InternVL on the BestShot Benchmark and a notable 68% improvement on the THUMOS14 Temporal Action Localization Benchmark (Jiang et al. 2014), while maintaining SOTA performance in general image classification and retrieval, as illustrated in Fig. 2[1].

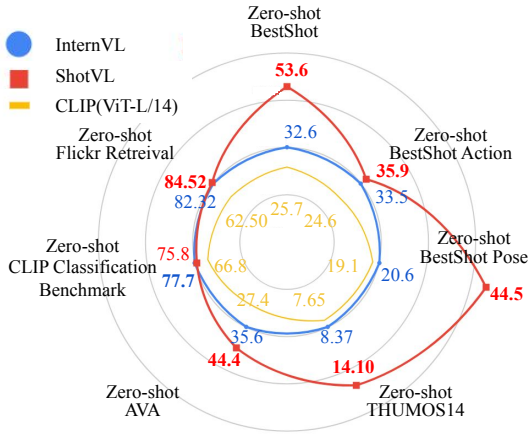


Figure 2: Zero-Shot evaluation on BestShot, Temporal Action Localization (THUMOS14), Action Classification (AVA), and CLIP Classification and Retrieval Benchmark.

## Related Work

### Datasets and Tasks

**Shot by Image aesthetics.** A straightforward approach to BestShot involves leveraging Image Quality Assessment (IQA) and Image Aesthetics Assessment (IAA) techniques, such as those used in the AVA dataset (Murray, Marchesotti, and Perronnin 2012), to select the highest-scoring frame. In practice, frame scoring is often integrated with query-based retrieval to enhance frame selection.

**Shot by Query.** Query-based moment retrieval includes Temporal Action Localization (TAL) and Moment Retrieval (MR) task. TAL datasets, such as THUMOS14, FineAction, MultiSports, and ActivityNet, are limited by fixed queries.

[1]We use both AVA Actions and AVA-Kinetics collectively as AVA.

Models trained on these datasets often lack zero-shot capabilities. MR datasets like Charades-STA, QVHighlight, HiRest (Zala et al. 2023), QueryD (Oncescu et al. 2021), and DiDeMo (Anne Hendricks et al. 2017) involve annotated queries but focus more on long temporal segments, emphasizing general video understanding rather than highlight frame retrieval in human-centric videos.

**Shot by Pose.** The use of pose descriptions for highlight retrieval remains underexplored. PoseScript (Delmas, Ginger and Weinzaepfel, Philippe and Lucas, Thomas and Moreno-Noguer, Francesc and Rogez, Grégory 2022) is the first large-scale pose description dataset, although it only annotates AMASS SMPL (Mahmood et al. 2019), rather than real-world images. Both Motion-X (Lin et al. 2024b) and our work utilize a PoseScript-like annotation method to describe poses in real-world videos. We additionally validated that this approach is highly effective for locating frames based on pose-queries in real-world videos. This general pose description approach satisfies certain highlight frame retrieval needs and is more user-friendly than dividing each action into dozens of specialized sub-actions.

**Shot by Video Question & Answer (Q&A).** Activitynet-RTL (Huang et al. 2024), Time-IT (Ren et al. 2024), and VTG-IT (Guo et al. 2024) transform moment retrieval annotations into video Q&As formats using GPT, without adding new localization data. Subsequently, they fine-tune LLMs on their proposed dataset, enabling LLMs to acquire moment retrieval capabilities in specific domains.

### Vision Language Models

**Image-based.** Although CLIP and its derivative methods (CoCa (Yu et al. 2022), InternVL) can directly perform frame retrieval in videos via language queries, they struggle with distinguishing subtle differences between adjacent frames and are less sensitive to long queries describing detailed information, such as fine-grained action stages and poses. Visual LLMs built on these base models, such as InternVL-Chat (Chen et al. 2023b, 2024b), face similar issues. ChatHuman (Lin et al. 2024a), which ensembles multiple models by LLMs, and PoseGPT (Lucas* et al. 2022), which regresses SMPL using LLMs, are more related to our work, as both focus on human-centric images.

**Video-based.** Although InternVideo performs well on supervised Temporal Action Localization tasks, demonstrating robust temporal understanding, it falls short in capturing fine-grained, frame-level details and in performing zero-shot retrieval within video, compared to CLIP and InternVL, which are trained only on image-text pairs. Current Video LLMs capable of retrieval within video, such as TimeChat, VTG-LLM, and LITA, often use image-based models, like CLIP, as the vision encoder rather than video-based models. In contrast, video LLMs built on video-based models (e.g. InternVideo), such as VideoChat (Li et al. 2024) and VideoGPT+ (Maaz et al. 2024), have yet to be validated for their localization capabilities in retrieval within videos.

## Benchmark and Training Data

In this section, we discuss the annotation methods for the BestShot Benchmark (Fig. 1) and training data for the proposed baseline model, ShotVL.

### Human-annotated Zero-Shot Benchmark

To construct the benchmark, we randomly selected 10% of the videos covering 60 action categories from THUMOS14 and FineAction for re-annotation. All ground truth in our benchmark is human-annotated, following a three-step pipeline:

**Selecting Potential Highlight Frames** identifies potential highlight frames from the sequences. Two groups of annotators score each frame on human expressiveness (considering action importance, pose extent, and facial expression, similar to Image Aesthetic Assessment but more objective), with scores ranging from 1 to 5. Frames with an average score above 4.5 are considered potential highlights.

**Writing detail queries for Highlight Frames** asks annotators to write a detailed description of the highlight frame that uniquely identifies it within the video. The descriptions are divided into three parts: (1) Visual Contents: Distinguish the main character from background characters, (2) Fine-Grained Action Stages: Describe key moments in actions in detail, e.g., "peak of jump" instead of "person is jumping", (3) Significant Human Poses: Annotate salient poses during actions using general descriptions. This is essential for professional actions, such as gymnastics, and complex actions, such as dance.

**Labeling Start-End Segments** defines the start and end frames in the video based on the text descriptions to ensure the accuracy of the ground truth query.

The Top@1 retrieval accuracy metric was used, where a prediction is considered correct if the predicted frame falls within any of the ground-truth intervals in the video. We divided 6,000 queries into three categories: Content, Action, and Pose, with 2,000 queries in each category. The "Full" metric combines all three query types into one, while "Action" focuses only on action queries, and "Pose" focuses only on pose queries. For "Pose", the ground-truth interval allows a tolerance of four frames around the key frame, while "Content" and "Action" use manually annotated intervals with an average of 12 frames.

### Hybrid Training Data and Annotation Pipeline

We followed three criteria to avoid benchmark leakage when constructing the training data: (1) The annotations we proposed were entirely generated by GPT or automatically synthesized, without human annotation[2]. (2) We did not sample frames or actions from the THUMOS14 or FineAction datasets. (3) Frames were randomly selected from the LAION-400M (Schuhmann et al. 2021) datasets, without focusing on the 60 action categories in the benchmark.

**ShotGPT4o Dataset:** We used GPT-4o to annotate image captions for a total of 600K frames. The datasets consists of: (1) 200K images with 800K captions, (2) another 150K images with 700K image Q&As, and (3) 250K frame

[2]Except for 2,000 human-written pose descriptions, crucial for aligning SMPLText, GPTPoseText, and Human-Written PoseText. These training data cause benchmark leakage, thus we presents a further ablation study related to this dataset in Table 4.

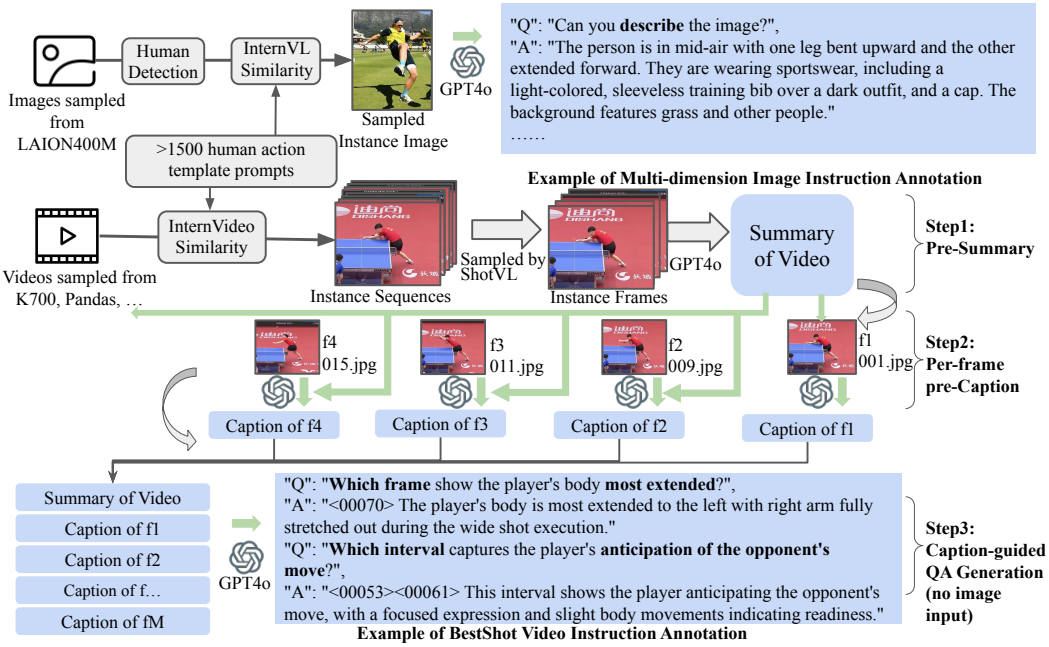


Figure 3: Annotation Pipeline of ShotGPT4o.

descriptions along with video Q&As from 10K videos, as well as 100K corresponding video Q&As, covering multiple tasks like frame retrieval, moment retrieval, dense captioning, video summary, and visual reasoning. The images are sampled from the LAION-400M dataset and videos are sampled from the K700 dataset, covering at least 1,500 action categories as shown in Fig. 3. Note that to avoid benchmark leakage from K700 to better evaluate zero-shot performance, the video Q&A annotations are not used in experiments shown in Tab. 3 and Tab. 4, but used in the video chat application shown in Fig. 8. Unlike prior work (Liu et al. 2023; Chen et al. 2023a, 2015) that focused solely on summary or fine-grained descriptions, we divided Q&As into four categories: summary, description, action and detailed action stage, and pose description. The annotation pipeline is illustrated in Fig. 3(a). For video Q&As, we also propose a novel annotation strategy, modified from ShareGPT4Video (Chen et al. 2024a), to further improve the accuracy of frame descriptions and frame or moment retrieval Q&As. Since this annotation strategy and its video Q&As are not crucial for training ShotVL, it serves as an initial attempt at frame retrieval using temporal cues. The effectiveness of this approach is shown in Fig. 3 and Fig. 8. More details will be provided in the extended version of the paper.

**Image-SMPLText Dataset:** BestShot Benchmark improvements were limited by the high noise level in GPT-4o's 375K pose descriptions. To address this, we introduce the Image-SMPLText Dataset, which adapts PoseScript for real-world videos to re-annotate over 13 public video datasets (Yi et al. 2023; von Marcard et al. 2018; Ionescu et al. 2014; Andriluka et al. 2018; Lin et al. 2023; Huang et al. 2022; Zhang et al. 2022; Cai et al. 2021; Kanazawa et al. 2019; Yang et al. 2023; Cheng et al. 2023). We enhanced PoseScript with body orientation to better suit real-world videos and generated 18.6M pose descriptions based on SMPL joint positions and orientations. Additionally, 2,000 samples were manually described to align SMPLText, pose descriptions in ShotGPT4o, and human-written descriptions. Although Motion-X also generates SMPLText, our focus is on real-world videos with accurate SMPL ground truth, unlike Motion-X's use of crawled data. As shown in Tab. 2, we only use the video datasets with accurate SMPL ground-truths which have been proven effective in SMPLer-X (Cai et al. 2024), as crawled video data often introduces noise that can degrade training performance. Fig. 4 shows an example of the difference between human-written, GPT-4o's generated descriptions, and Image-SMPLText.

**Human-written:** This person is facing the camera, using his left foot as a fulcrum, kicking his right foot forward, with his right foot parallel to the ground, his hands raised upwards, and his body facing the seven o'clock direction.

**GPT4o:** The person is in a seated position on the ground with legs extended forward and knees slightly bent. Both arms are bent at the elbows, with the forearms raised and hands open, palms facing outward.

**ImageSMPLText:** The subject is bent forward, the right upper arm is parallel to the ground, the right elbow is in the back of the left elbow, the right elbow is bent slightly and the right hand is higher than the right shoulder while the right hand is located behind the left hand with the right hand spread far apart from the left hand with the left elbow bent, the left hand is above the left shoulder while the right knee is bent slightly with the right foot in front of the left foot, the knees are about shoulder width apart. The left knee is fully bent while the left calf is parallel to the ground, the feet are spread far apart, the left foot is behind the torso.

Figure 4: Difference of pose descriptions among Human-Written, GPT-4o, and Image-SMPLText.

| Motion-X | | | Ours | | |
|---|---|---|---|---|---|
| **Dataset** | **#Clip** | **#Frame** | **Dataset** | **#Clip** | **#Frame** |
| Aist++ | 1470 | 340K | BEDLAM | 10000 | 951K |
| Animation*2 | 329 | 38K | SynBody | 6K | 633K |
| Dance*2 | 163 | 36K | InstaVariety | 14K | 2.1M |
| Egobody | 980 | 438K | GTA-H II | 10K | 1.8M |
| Fitness*2 | 16.7K | 358K | EgoBody | 125 | 935k |
| Game*2 | 10.2K | 1.1M | RICH | 496 | 243K |
| GRAB*1 | 1.3K | 406k | UBody | 836 | 683k |
| HAA500 | 5.2K | 311K | PoseTrack | 1.3K | 28K |
| Humanml*1 | 26K | 3.5M | BEHAVE | 299 | 44K |
| Humman | 744 | 104K | H3.6M | 840 | 312K |
| Idea400 | 12.5K | 2.5M | 3DPW | 61 | 22.7K |
| kungfu*2 | 1040 | 257K | DNA-R | 24K | 5M |
| perform*2 | 475 | 102K | TalkShow | 984 | 3.3M |

Table 2: Components of Public Datasets. *1 indicates that the dataset contains no images, *2 indicates that the dataset is sourced from the Internet and may have no accurate SMPL. In comparison, the Motion-X dataset does not entirely consist of images, and a significant portion is sourced from videos, which can lead to issues such as inaccurate keypoint predictions and imprecise pose descriptions. In contrast, we only use publicly available video datasets with accurate SMPL ground truth. For comparison, PoseScript annotates only the AMASS dataset, which lacks real-world images.

## Analysis & Limitation of Benchmark & Data

**Zero-Shot BestShot Benchmark.** Tab. 1 presents the details of the BestShot Benchmark, which is collected from 112K frames of THUMOS14 and 2M frames of FineAction, covering 400 clips of THUMOS14 and 500 videos of FineAction. Fig. 1 shows two query samples. To validate the benchmark's accuracy, a separate group of annotators located keyframes in the videos based on the queries. The average human retrieval rate was 86%, significantly surpassing current SOTA methods.

**ShotGPT4o.** Manual validation of 500 Q&As revealed that only 70% of pose descriptions are accurate, with GPT-4o often confusing left and right. This lower accuracy accounts for the poor results when training solely on GPT-4o's pose descriptions. Additionally, although 90% of action descriptions are correct, only 22% are fine-grained. For instance, 78% of throwing actions are described in broad terms as "extending the body during a throw", while only 22% capture detailed stages like "about to release the ball", "just released the ball", or "at maximum power". This explains why the BestShot Benchmark's action dimension improves by only 10% with GPT-4o data.

**Image-SMPLText.** There are still some problems to address. The first is the lack of temporal information, which is partly solved in PoseFix (Delmas et al. 2024) and MotionScript (Yazdian et al. 2023). They re-wrote the pose-code in PoseScript to generate a caption of two frames and a few sequences. We also generated more than 18.6M sequence descriptions in this way, but they have not yet been used in training. The second issue is the insufficient pose-code related to real-world videos. Although we have addressed the lack of a global orientation code, pose information related to translation and velocities has not yet been completed. The third problem is that the unseen body parts are also described in detail since the pose caption is automatically generated from ground truth SMPL. Specially, we assume that keypoints with lower confidence were occluded and describe them as "... is occluded" or choose not to describe them. However, experiments show that confidence cannot accurately reflect occlusion in most cases. Given that the model can infer the pose of an occluded body part, we ultimately decided not to introduce the concept of occlusion in the final implementation.

## Method

In this section, we present the proposed baseline model, ShotVL (Fig. 5). Given a video and text query, ShotVL is able to locate the most relevant frames or moments. To address the challenges of the BestShot task, we trained ShotVL following the fine-tuning stage of the InternVL-Base model. The key designs used in ShotVL are listed as follows:

**Base Model Selection.** Although InternVideo excels at Supervised Temporal Action Localization tasks like THUMOS14 and FineAction, the BestShot task demands fine-grained image understanding and precise localization of very short sequences. Therefore, InternVideo, pre-trained only on video-text pairs, underperforms compared to the smaller CLIP model. On the other hand, InternVL significantly outperforms CLIP in Zero-Shot BestShot tasks, making it the preferred baseline model for our approach.

**Furthest Point Sampling.** We applied furthest point sampling (FPS) on the 18.6M Image-SMPLText samples to create 10 different 1% and 2% subsets from varying start frames to avoid the impact of excessive similarity between adjacent frames. First, we randomly selected a frame as the start frame and calculated the mean per-joint position error (MPJPE) after orientation alignment with other frames. The frame with the largest MPJPE would be selected as a new start frame. This process was repeated until the required number of frames was sampled from the entire dataset.

**Customized Data and Data Ratios.** Due to the significant differences between Image-SMPLText and the original InternVL training data, training only on SMPLText weakens the pre-trained model's performance on general image tasks. To address this, we added the COCO dataset (Chen et al. 2015), a high-quality human-written image caption dataset, and adjusted data ratios to preserve generalization capabilities. Additionally, we replaced some Image-SMPLText descriptions with GPT-4o-generated or human-written captions to bridge the style gap between the dataset descriptions and general queries.

**Inference Pipeline.** During inference, frame-by-frame similarities with the queries are computed first. For the BestShot task, the best frame is obtained by simply applying argmax or using non-maximum suppression (NMS) to select the predicted frames. For moment retrieval or temporal action localization task, we follow T3AL (T=0) (Liberatori et al. 2024), which first uses the mean of video vision feature to match a pseudo-label from the list of possible actions, and then segment the sequences based on their similarities.

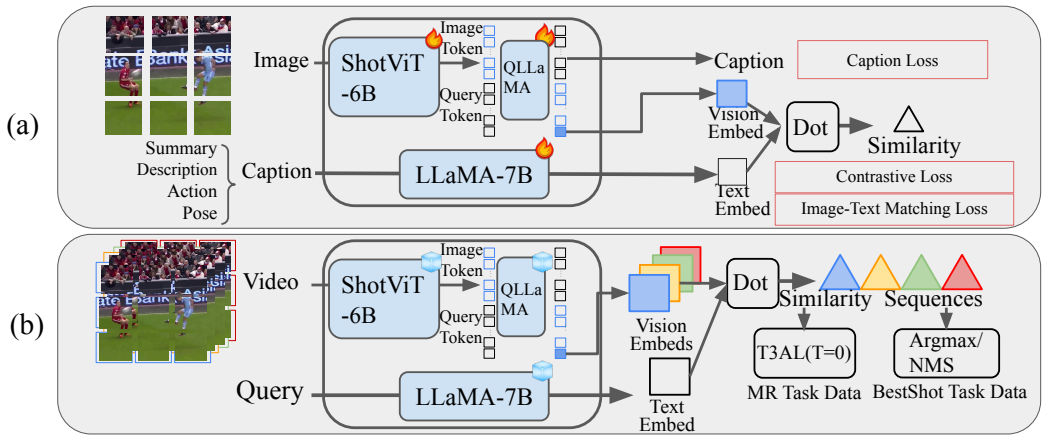


Figure 5: Training and inference pipeline of ShotVL. (a) Training. (b) Inference of BestShot and Moment Retrieval.

| method | size | Zero-Shot BestShot Full (top1) | Zero-Shot BestShot Action (top1) | Zero-Shot BestShot Pose (top1) | Zero-Shot TAL THUMOS14 (avg mAP) | Zero-Shot AC AVA mean (acc1/5) | Zero-Shot CLIP Benchmark flickr,I2T, (top1) | Zero-Shot CLIP Benchmark flickr,T2I (top1) | Zero-Shot CLIP Benchmark avg(C) (top1) |
|---|---|---|---|---|---|---|---|---|---|
| (a1)CLIP ViT-B/32 | 188M | 24.6 | 19.1 | 7.6 | - | 21.00 | 73.75 | 55.38 | 53.78 |
| (a2)CLIP ViT-L/14 | 406M | 25.7 | 24.6 | 19.1 | 7.65 | 27.39 | 81.45 | 62.50 | 66.75 |
| (a3)LongCLIP | 406M | 31.3 | 30.1 | 21.2 | 7.41 | 35.70 | 86.35 | 74.28 | 68.95 |
| (a4)InternVL | 14B | 32.6 | 33.5 | 20.6 | 8.37 | 35.57 | 92.40 | 82.32 | **77.70** |
| (b)InternVideo1 | 1.3B | 21.6 | 23.8 | 17.7 | 4.00 | * | × | × | × |
| (c1)VTG-LLM | 8.4B | 6.2 | 7.6 | 5.1 | 0.66 | × | × | × | × |
| (c2)LITA | 13.2B | 12.6 | 12.2 | 14.1 | 0.17 | × | × | × | × |
| (d)**ShotVL** | 14B | **53.6** | **35.9** | **44.5** | **14.10** | **44.40** | **94.50** | **84.52** | 75.82 |

Table 3: Quantitative comparison of SOTA models on zero-shot BestShot Benchmark, THUMOS14 Temporal Action Localization (TAL), AVA Classification (AC), CLIP Retrieval Benchmark (only Flickr (Young et al. 2014) evaluated; COCO excluded due to benchmark leakage during training; I2T and T2I denote image-to-text and text-to-image), and CLIP Classification Benchmark (average of 24 tasks). TAL testing uses the T3AL(T=0) method for the base model. *InternVideo's zero-shot AVA-Kinetics (Li et al. 2020) result is not evaluated due to prior training on Kinetics. (a) SOTA base models with single-frame input, (b) SOTA base models with multi-frame input, (c) SOTA video LLMs with multi-frame input and segment output. For frame prediction, we use the middle frame of the predicted interval, as it yields the highest score compared to other frames. We evaluated top1 accuracy in BestShot, the mean of mAP under iou=0.3,0.4,0.5,0.6,0.7 in THUMOS14, the mean of top1 and top5 accuracy in AVA, and top1 accuracy in the CLIP Benchmark.

## Experiment

Quantitative comparison of SOTA models on zero-shot BestShot Benchmark, THUMOS14 Temporal Action Localization, AVA Classification, CLIP Retrieval Benchmark (only Flickr evaluated; COCO excluded due to benchmark leak during our training; I2T and T2I denote image-to-text and text-to-image), and CLIP Classification Benchmark (average of 20 tasks). TAL testing uses the T3AL(T=0) method (Liberatori et al. 2024) for the base model.

**Implementation of ShotVL.** ShotVL follows the fine-tuning pipeline of InternVL with our specially designed datasets selection and ratios. We used InternVL 14B as the base model. The datasets are divided into 3 parts: SMPL-Text, ShotGPT4o, and General, with a ratio of 1:5:5, which ensures a stable balance between the BestShot task and general retrieval/classification tasks. The ShotVL model was trained for 2,000 iterations, with batch size 1,536 and a learning rate of 1e-5, on 24 A100 GPUs for 20 hours.

**Evaluation of ShotVL.** We evaluated several SOTA models and ShotVL for zero-shot frame localization (BestShot), temporal localization (THUMOS14 TAL), action classification (AVA), and CLIP general retrieval and classification benchmarks. As shown in Tab. 3, benchmark performance improves with model size. CLIP struggles with fine-grained long-text retrieval on BestShot due to limited training text length. While LongCLIP (Zhang et al. 2024) with positional encoding interpolation offer partial solutions, the lack of long-text training data limits progress. InternVideo, despite large-scale video-text training, loses frame-level spatial details, affecting performance. InternVL excels in fine-grained understanding without CLIP's long-text constraints, showing significant BestShot improvements with simple fine-tuning. ShotVL outperforms InternVL across all metrics, matching its performance on general CLIP tasks.

**Ablation Study.** Ablation experiments are conducted in four groups (Tab. 4). From (a1) to (a2), InternVL was

| method | Finetuned Data | | | | | Zero-Shot BestShot | | | Zero-Shot TAL | Zero-Shot AC | Zero-Shot CLIP Benchmark | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Human Pose | GPT Pose | SMPL Pose | Shot GPT4o | General | Full | Action | Pose | THUMOS14 | AVA | flickr, I2T | flickr, T2I | avg(C) |
| (a1)InternVL | N | N | N | N | N | 32.6 | 33.5 | 20.6 | 8.37 | 35.57 | 92.40 | 82.32 | 77.70 |
| (a2)InternVL-ft | Y | N | N | N | N | 38.3 | 32.7 | 29.5 | 9.55 | **47.31** | 92.05 | 81.87 | **77.90** |
| (b1)ShotVL-PoseS | N | N | **Y** | N | N | 35.6 | 24.1 | 30.3 | 5.99 | 30.40 | 67.00 | 51.28 | 73.52 |
| (b2)ShotVL-PoseSH | **Y** | N | **Y** | N | N | 47.0 | 23.0 | 42.6 | 5.65 | 42.45 | 71.70 | 56.46 | 74.45 |
| (b3)ShotVL-Pose | **Y** | **Y** | **Y** | N | N | 49.1 | 24.3 | 43.6 | 5.53 | 35.87 | 73.65 | 57.82 | 74.30 |
| (c)ShotVL-G | **Y** | **Y** | **Y** | N | LAION | 46.6 | 31.8 | 43.4 | 8.32 | 40.71 | 89.25 | 76.38 | 77.40 |
| | **Y** | **Y** | **Y** | N | COCO | 52.0 | 33.8 | **45.1** | 10.30 | 45.03 | 94.25 | 84.14 | 76.07 |
| (d)ShotVL-Full | **Y** | **Y** | **Y** | **Y** | COCO | **53.6** | **35.9** | 44.5 | **14.10** | 44.40 | **94.50** | **84.52** | 75.82 |

Table 4: Ablation Study on Finetuned Data and Model Performance for InternVL and ShotVL (ShotVL refers to the model trained on our proposed ShotGPT4o and Image-SMPLText datasets). Human/GPT Pose (Human-Written and GPT-generated pose descriptions), SMPL Pose (our Image-SMPLText dataset), General (image captioning datasets such as LAION-400M and COCO captions), Zero-Shot TAL (Zero-Shot Temporal Action Localization), and Zero-Shot AC (Zero-Shot Action Classification). Metrics are the same as Tab. 4.

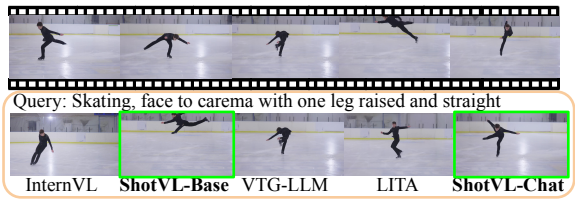


Figure 6: Frame Retrieval out of BestShot Benchmark (Videos of Skating).

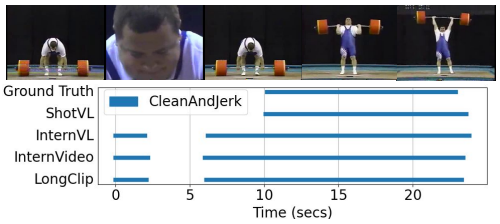


Figure 7: An example of qualitative comparisons of Zero-Shot TAL.

fine-tuned using 2,000 human-annotated pose descriptions. From (b1) to (b3), we assessed the effectiveness of Image-SMPLText and the importance of aligning the three types of pose descriptions. However, the introduction of Image-SMPLText led to a significant decline in general retrieval performance. Therefore, starting from (c), we incorporated general image-text data into the training to preserve both pose retrieval and general retrieval capabilities. Finally, (d) validated the effectiveness of ShotGPT4o.

**Generalization of BestShot.** For queries and scenes not covered in the evaluated benchmarks, Fig. 6 demonstrated ShotVL's strong generalization capability.

**Zero-Shot Temporal Action Localization.** CLIP, InternVL, and ShotVL can be directly applied to zero-shot temporal action localization using the T3AL ($T = 0$) method. The comparisons are shown in Fig. 2 and Fig. 7.

**Video Chat.** ShotVL can be further integrated with a

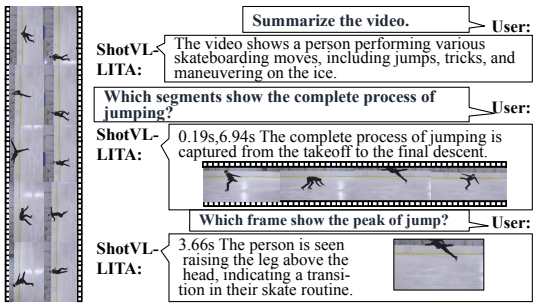


Figure 8: An example of video chat application.

Video LLM to enable video question answering, moment retrieval, and frame retrieval simultaneously as shown in Fig. 8. We adapt LITA as the baseline method to train our Video Chat model. The modifications include: (1) replacing the encoder with ShotVL, (2) pre-training and fine-tuning on image Q&As following the LLaVA approach, (3) adding a frame token into the LLM, (4) fine-tuning on multiple video instruction datasets, including ShotGPT4o.

## Conclusion

We collected the BestShot Benchmark, a highlight-frame-retrieval-in-video benchmark that includes 6,000 queries, with manually identified highlight frames, detailed descriptions, and matched temporal segments. We further proposed two datasets tailored for the zero-shot BestShot task: ShotGPT4o and Image-SMPLText. These datasets have been proven effective on zero-shot BestShot, AVA classification, and THUMOS14 action localization benchmarks.

Methodologically, we evaluated several SOTA models on the BestShot task and proposed ShotVL as a robust solution. ShotVL shows superior zero-shot performance on the BestShot task, yet it is still limited in retrieving actions that require strong temporal correlations. A promising future direction is to extend ShotVL by integrating it with Video LLMs and collect a large-scale video dataset for frame or moment retrieval to train such a model.

## References


Andriluka, M.; Iqbal, U.; Insafutdinov, E.; Pishchulin, L.; Milan, A.; Gall, J.; and Schiele, B. 2018. PoseTrack: A Benchmark for Human Pose Estimation and Tracking. arXiv:1710.10000.

Anne Hendricks, L.; Wang, O.; Shechtman, E.; Sivic, J.; Darrell, T.; and Russell, B. 2017. Localizing moments in video with natural language. In *Proceedings of the IEEE international conference on computer vision*, 5803–5812.

Cai, Z.; Yin, W.; Zeng, A.; Wei, C.; Sun, Q.; Wang, Y.; Pang, H. E.; Mei, H.; Zhang, M.; Zhang, L.; Loy, C. C.; Yang, L.; and Liu, Z. 2024. SMPLer-X: Scaling Up Expressive Human Pose and Shape Estimation. arXiv:2309.17448.

Cai, Z.; Zhang, M.; Ren, J.; Wei, C.; Ren, D.; Lin, Z.; Zhao, H.; Yang, L.; and Liu, Z. 2021. Playing for 3D human recovery. *arXiv preprint arXiv:2110.07588*.

Chen, L.; Li, J.; Dong, X.; Zhang, P.; He, C.; Wang, J.; Zhao, F.; and Lin, D. 2023a. ShareGPT4V: Improving Large Multi-Modal Models with Better Captions. *arXiv preprint arXiv:2311.12793*.

Chen, L.; Wei, X.; Li, J.; Dong, X.; Zhang, P.; Zang, Y.; Chen, Z.; Duan, H.; Lin, B.; Tang, Z.; Yuan, L.; Qiao, Y.; Lin, D.; Zhao, F.; and Wang, J. 2024a. ShareGPT4Video: Improving Video Understanding and Generation with Better Captions. *arXiv preprint arXiv:2406.04325*.

Chen, X.; Fang, H.; Lin, T.-Y.; Vedantam, R.; Gupta, S.; Dollar, P.; and Zitnick, C. L. 2015. Microsoft COCO Captions: Data Collection and Evaluation Server. arXiv:1504.00325.

Chen, Z.; Wang, W.; Tian, H.; Ye, S.; Gao, Z.; Cui, E.; Tong, W.; Hu, K.; Luo, J.; Ma, Z.; Ma, J.; Wang, J.; Dong, X.; Yan, H.; Guo, H.; He, C.; Shi, B.; Jin, Z.; Xu, C.; Wang, B.; Wei, X.; Li, W.; Zhang, W.; Zhang, B.; Cai, P.; Wen, L.; Yan, X.; Dou, M.; Lu, L.; Zhu, X.; Lu, T.; Lin, D.; Qiao, Y.; Dai, J.; and Wang, W. 2024b. How Far Are We to GPT-4V? Closing the Gap to Commercial Multimodal Models with Open-Source Suites. arXiv:2404.16821.

Chen, Z.; Wu, J.; Wang, W.; Su, W.; Chen, G.; Xing, S.; Zhong, M.; Zhang, Q.; Zhu, X.; Lu, L.; Li, B.; Luo, P.; Lu, T.; Qiao, Y.; and Dai, J. 2023b. InternVL: Scaling up Vision Foundation Models and Aligning for Generic Visual-Linguistic Tasks. *arXiv preprint arXiv:2312.14238*.

Cheng, W.; Chen, R.; Yin, W.; Fan, S.; Chen, K.; He, H.; Luo, H.; Cai, Z.; Wang, J.; Gao, Y.; Yu, Z.; Lin, Z.; Ren, D.; Yang, L.; Liu, Z.; Loy, C. C.; Qian, C.; Wu, W.; Lin, D.; Dai, B.; and Lin, K.-Y. 2023. DNA-Rendering: A Diverse Neural Actor Repository for High-Fidelity Human-centric Rendering. *arXiv preprint*, arXiv:2307.10173.

Delmas, G.; Weinzaepfel, P.; Moreno-Noguer, F.; and Rogez, G. 2024. PoseFix: Correcting 3D Human Poses with Natural Language. arXiv:2309.08480.

Delmas, Ginger and Weinzaepfel, Philippe and Lucas, Thomas and Moreno-Noguer, Francesc and Rogez, Grégory. 2022. PoseScript: 3D Human Poses from Natural Language. In *ECCV*.

Guo, Y.; Liu, J.; Li, M.; Tang, X.; Chen, X.; and Zhao, B. 2024. VTG-LLM: Integrating Timestamp Knowledge into Video LLMs for Enhanced Video Temporal Grounding. *arXiv preprint arXiv:2405.13382*.

Huang, C.-H. P.; Yi, H.; Höschle, M.; Safroshkin, M.; Alexiadis, T.; Polikovsky, S.; Scharstein, D.; and Black, M. J. 2022. Capturing and Inferring Dense Full-Body Human-Scene Contact. In *Proceedings IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 13274–13285.

Huang, D.-A.; Liao, S.; Radhakrishnan, S.; Yin, H.; Molchanov, P.; Yu, Z.; and Kautz, J. 2024. Lita: Language instructed temporal-localization assistant. *arXiv preprint arXiv:2403.19046*.

Ionescu, C.; Papava, D.; Olaru, V.; and Sminchisescu, C. 2014. Human3.6M: Large Scale Datasets and Predictive Methods for 3D Human Sensing in Natural Environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(7): 1325–1339.

Jiang, Y.-G.; Liu, J.; Roshan Zamir, A.; Toderici, G.; Laptev, I.; Shah, M.; and Sukthankar, R. 2014. THUMOS Challenge: Action Recognition with a Large Number of Classes. http://crcv.ucf.edu/THUMOS14/.

Kanazawa, A.; Zhang, J. Y.; Felsen, P.; and Malik, J. 2019. Learning 3D Human Dynamics from Video. In *Computer Vision and Pattern Recognition (CVPR)*.

Li, A.; Thotakuri, M.; Ross, D. A.; Carreira, J.; Vostrikov, A.; and Zisserman, A. 2020. The AVA-Kinetics Localized Human Actions Video Dataset. arXiv:2005.00214.

Li, K.; He, Y.; Wang, Y.; Li, Y.; Wang, W.; Luo, P.; Wang, Y.; Wang, L.; and Qiao, Y. 2024. VideoChat: Chat-Centric Video Understanding. arXiv:2305.06355.

Liberatori, B.; Conti, A.; Rota, P.; Wang, Y.; and Ricci, E. 2024. Test-Time Zero-Shot Temporal Action Localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 18720–18729.

Lin, J.; Feng, Y.; Liu, W.; and Black, M. J. 2024a. ChatHuman: Language-driven 3D Human Understanding with Retrieval-Augmented Tool Reasoning. arXiv:2405.04533.

Lin, J.; Zeng, A.; Lu, S.; Cai, Y.; Zhang, R.; Wang, H.; and Zhang, L. 2024b. Motion-X: A Large-scale 3D Expressive Whole-body Human Motion Dataset. arXiv:2307.00818.

Lin, J.; Zeng, A.; Wang, H.; Zhang, L.; and Li, Y. 2023. One-Stage 3D Whole-Body Mesh Recovery with Component Aware Transformer. *CVPR*.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023. Visual Instruction Tuning. In *NeurIPS*.

Lucas*, T.; Baradel*, F.; Weinzaepfel, P.; and Rogez, G. 2022. PoseGPT: Quantization-based 3D Human Motion Generation and Forecasting. In *European Conference on Computer Vision (ECCV)*.

Maaz, M.; Rasheed, H.; Khan, S.; and Khan, F. 2024. VideoGPT+: Integrating Image and Video Encoders for Enhanced Video Understanding. arXiv:2406.09418.

Mahmood, N.; Ghorbani, N.; Troje, N. F.; Pons-Moll, G.; and Black, M. J. 2019. AMASS: Archive of Motion Capture as Surface Shapes. In *International Conference on Computer Vision*, 5442–5451.

Murray, N.; Marchesotti, L.; and Perronnin, F. 2012. AVA: A large-scale database for aesthetic visual analysis. In *2012 IEEE conference on computer vision and pattern recognition*, 2408–2415. IEEE.

Oncescu, A.-M.; Henriques, J. F.; Liu, Y.; Zisserman, A.; and Albanie, S. 2021. Queryd: A video dataset with high-quality text and audio narrations. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2265–2269. IEEE.

Ren, S.; Yao, L.; Li, S.; Sun, X.; and Hou, L. 2024. Timechat: A time-sensitive multimodal large language model for long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14313–14323.

Schuhmann, C.; Vencu, R.; Beaumont, R.; Kaczmarczyk, R.; Mullis, C.; Katta, A.; Coombes, T.; Jitsev, J.; and Komatsuzaki, A. 2021. LAION-400M: Open Dataset of CLIP-Filtered 400 Million Image-Text Pairs. arXiv:2111.02114.

von Marcard, T.; Henschel, R.; Black, M.; Rosenhahn, B.; and Pons-Moll, G. 2018. Recovering Accurate 3D Human Pose in The Wild Using IMUs and a Moving Camera. In *European Conference on Computer Vision (ECCV)*.

Wang, Y.; Li, K.; Li, Y.; He, Y.; Huang, B.; Zhao, Z.; Zhang, H.; Xu, J.; Liu, Y.; Wang, Z.; Xing, S.; Chen, G.; Pan, J.; Yu, J.; Wang, Y.; Wang, L.; and Qiao, Y. 2022. InternVideo: General Video Foundation Models via Generative and Discriminative Learning. *arXiv preprint arXiv:2212.03191*.

Yang, Z.; Cai, Z.; Mei, H.; Liu, S.; Chen, Z.; Xiao, W.; Wei, Y.; Qing, Z.; Wei, C.; Dai, B.; Wu, W.; Qian, C.; Lin, D.; Liu, Z.; and Yang, L. 2023. SynBody: Synthetic Dataset with Layered Human Models for 3D Human Perception and Modeling. arXiv:2303.17368.

Yazdian, P. J.; Liu, E.; Cheng, L.; and Lim, A. 2023. MotionScript: Natural Language Descriptions for Expressive 3D Human Motions. arXiv:2312.12634.

Yi, H.; Liang, H.; Liu, Y.; Cao, Q.; Wen, Y.; Bolkart, T.; Tao, D.; and Black, M. J. 2023. Generating Holistic 3D Human Motion from Speech. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 469–480.

Young, P.; Lai, A.; Hodosh, M.; and Hockenmaier, J. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2: 67–78.

Yu, J.; Wang, Z.; Vasudevan, V.; Yeung, L.; Seyedhosseini, M.; and Wu, Y. 2022. CoCa: Contrastive Captioners are Image-Text Foundation Models. arXiv:2205.01917.

Zala, A.; Cho, J.; Kottur, S.; Chen, X.; Oguz, B.; Mehdad, Y.; and Bansal, M. 2023. Hierarchical video-moment retrieval and step-captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 23056–23065.

Zhang, B.; Zhang, P.; Dong, X.; Zang, Y.; and Wang, J. 2024. Long-CLIP: Unlocking the Long-Text Capability of CLIP. *arXiv preprint arXiv:2403.15378*.

Zhang, S.; Ma, Q.; Zhang, Y.; Qian, Z.; Kwon, T.; Pollefeys, M.; Bogo, F.; and Tang, S. 2022. Egobody: Human body shape and motion of interacting people from head-mounted devices. In *European Conference on Computer Vision*, 180–200. Springer.