HAPACT: A Benchmark For Human-Centric Physical Impact Localization in Movies

Youngrae Kim[2], Yejin Jang[3], Eunho Kim[1], Youjin Sung[1], Kun-Woo Song[1], Sangpil Kim[4], Sang Ho Yoon[1,2,3]

[1] Graduate School of Culture Technology, KAIST, South Korea
[2] Graduate School of Metaverse, KAIST, South Korea
[3] Department of AX, KAIST College of AI, South Korea
[4] Department of Artificial Intelligence, Korea University, South Korea

Abstract

Physical reasoning about humans in videos has largely focused on everyday interactions, where contact, pressure, pose, and motion support embodied perception and contact-aware applications. Still, short and forceful properties like impact events remain relatively underexplored, despite their relevance to broader understanding of how abrupt physical events affect the human body. To fill this gap, we introduce human-centric impact localization, a new task that localizes target physical impacts temporally within a video and spatially on the human body. We further present **HAPACT**, a HumAn-centric imPACT dataset and benchmark built from movie clips, containing over 2.8K video-query pairs with more than 20K fine-grained human-annotated frames. We built HAPACT around viewer-grounded annotation, text-conditioned localization, and supervision at frame-level temporal and SMPL-X-based spatial granularity. To demonstrate the effectiveness of our benchmark, we conduct extensive evaluations of existing methods, employing video moment retrieval models for temporal localization and human-scene contact models for spatial localization. For the spatial task, we propose a text-conditioned baseline by reformulating localization as query-guided body-region selection. Benchmark results show that physical impact localization remains challenging for existing methods, motivating HAPACT as a dedicated benchmark to support physical impact reasoning in videos.

1 Introduction

Physical AI and multimodal foundation models increasingly require human-centered physical reasoning from videos, going beyond action recognition to understand how physical events affect the human body. Recent body-centric physical interaction datasets examine how the body interacts with surrounding scenes [11, 24, 27, 64], objects [15, 32, 44, 60, 68, 71], and other people [35, 69, 72]. They mainly focus on everyday user-driven interaction regimes, providing supervision for physically grounded cues such as contact, pressure, pose, and motion. Still, externally initiated interactions such as bodily impacts have remained overlooked.

Impact events offer a complementary axis for human-centric physical understanding. Unlike sustained or continuous interactions, impacts are short and forceful events in which the human body is suddenly subjected to external force [17]. Understanding such events requires reasoning about when and how external elements affect the body, as well as where those effects occur. Unlike common physical cues that arise from user-initiated interactions, impacts often arise from external agents or the environment, making them difficult to capture with active interaction cues alone. This shifts the focus from

40th Conference on Neural Information Processing Systems (NeurIPS 2026). Track on Evaluations and Datasets.

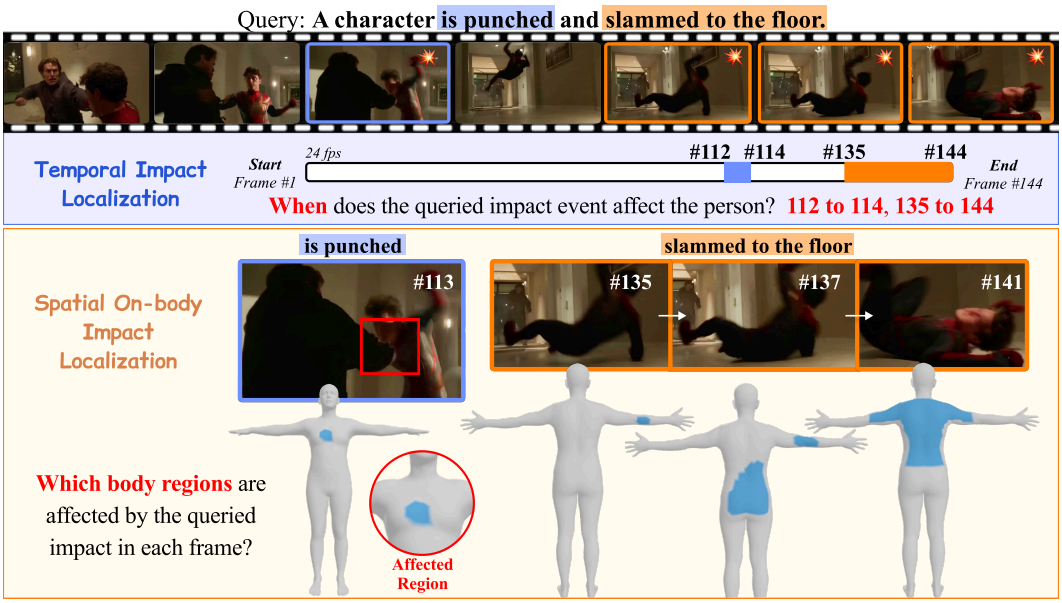


Figure 1: **Human-Centric Impact Localization.** Given a video and a text query describing a target impact event, the task requires localizing when the queried impact affects the person and which body regions are affected in each frame.

sustained interaction modeling to a broader understanding of how both user-initiated and externally initiated physical events affect the human body.

Action-oriented movies provide a rich domain for understanding impact, as they frequently depict bodily impacts by direct collisions and environmental forces in realistic scenes. Beyond their diverse impact depictions, movies also offer rich scenes and narrative context for video understanding, as leveraged in prior movie-based benchmarks [21, 58]. This context requires models to identify the specific impact described by a text query, rather than simply detect any events in the video. Reasoning about physical events within their video context can also support applications in immersive media such as haptic authoring and automated haptic experience generation [33, 34, 39].

This motivates our formulation of **Human-Centric Impact Localization**, a new task that localizes a target physical impact temporally within a video and spatially on the human body. Building on this task, we introduce **HAPACT**, a HumAn-centric imPACT dataset and benchmark for impact localization in movie clips. The dataset contains over 2.8K video-query pairs with more than 20K fine-grained human-annotated frames. To make the annotation procedure tractable, HAPACT uses a semi-automated pipeline that first generates coarse candidates with a vision-language model, then annotators create fine-grained labels using an interactive web-based interface. HAPACT is designed around three key properties:

- **Viewer-grounded annotation.** All annotations are labeled for impacts identified by human viewers. This allows HAPACT to better align its annotations with natural human interpretation of impact events in the video context and cinematic cues.
- **Text-conditioned localization.** Each textual query specifies a target impact event, and the task requires localizing the target event rather than the whole impact in the scene. This query-conditioned formulation supports event-level reasoning in complex movie scenes.
- **Fine-grained supervision.** HAPACT provides frame-level temporal annotations at 24 FPS and SMPL-X [51] mesh annotations for fine-grained spatio-temporal evaluation.

As a benchmark, HAPACT enables multi-axis evaluation across temporal and spatial axes. We evaluate temporal localization with existing video moment retrieval models and spatial localization with human-scene contact models. Since existing human-scene contact models are not conditioned on textual input, we further formulate on-body impact localization as query-guided body region selection and instantiate it by fine-tuning a multi-modal large language model (MLLM). While this

region selection baseline outperforms contact-based models [27, 41, 64], it remains limited to coarse body regions rather than fine-grained estimations on the body mesh. By bringing bodily impacts into physical reasoning, HAPACT broadens human physical understanding beyond user-intended interaction, providing a benchmark foundation for richer reasoning about dynamic environmental events and their consequences.

To summarize, the contributions are as follows:

- We formulate human-centric impact localization, a new task that identifies each queried physical impact with its corresponding video frames and predicts the affected region of the human body.
- HAPACT is the first human-centric video benchmark for physical impact events, providing viewer-grounded, text-conditioned annotations with frame-level SMPL-X labels.
- We evaluate representative models for temporal localization and spatial on-body localization, employing video moment retrieval models and human-scene contact models. We further introduce an MLLM baseline for query-guided region selection.

2 Related Work

Body-centric Physical Interaction Datasets. Recent body-centric datasets have expanded from modeling humans in physical scenes to richer cues about how the body interacts with its surroundings. Hassan et al. [24] was an early effort to use contact and penetration constraints to recover physically plausible human poses in real environments, while subsequent studies [11, 27, 56, 64] extended this direction toward dense body representations, including SMPL-X [51], and contact estimation in in-the-wild images. Beyond scene constraints, later datasets broaden this perspective to other interaction domains, spanning from human-object interaction [5, 15, 28, 44, 60, 68, 70, 71] to human-human interaction [35, 69, 72] between multiple bodies. Recent datasets further enrich body-centric physical cues like hand-pressure estimation, but remain largely focused on hand interactions [19, 55, 78]. While these datasets provide rich supervision for body-centric physical cues, they primarily cover everyday or user-initiated interactions, leaving external, environment-initiated physical events less represented. HAPACT complements this line by introducing impact as a physical cue in movie videos, capturing bodily effects from short, abrupt external force.

Video Moment Retrieval and Spatio-temporal Grounding Benchmarks. Video localization benchmarks [1, 16, 36, 38, 53, 75] have studied temporal event understanding not only through action localization, but also through moment retrieval guided by language queries across diverse video domains. Recent benchmarks further expand this language-guided moment retrieval along several axes, including corpus-level retrieval [10, 37], egocentric [20, 52], long [14, 57, 74] video domains, and multiple relevant moments [6, 73].

Beyond temporal localization, spatio-temporal video grounding (STVG) localizes targets specified by language jointly across video frames and image space. Existing STVG benchmarks [62, 77] provide foundations for joint spatial and temporal localization. Here, their spatial targets are usually defined in image coordinates, using boxes or tubes around target instances. HAPACT follows this formulation but shifts the spatial target from image-level anchors to frame-level body regions. This representation captures the effect of a queried physical event on the human body.

3 HAPACT

HAPACT is a movie-based dataset for human-centric impact understanding in videos. We target short and forceful events in which the human body is affected by collisions, falls, object interactions, or environmental forces. The dataset is built from over 5 hours of high-resolution videos spanning 212 movies, and contains over 2.8K video-query pairs with clips ranging from 6 to 8 seconds. The following sections detail the data collection and annotation procedures.

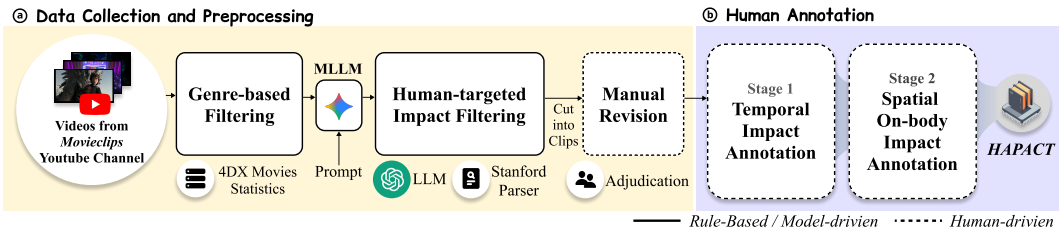


Figure 2: **Overview of the HAPACT Construction Pipeline.** Videos from the *Movieclips* YouTube channel are filtered by genre priors, processed with Gemini 2.5 Pro [13] to identify impacts and generate captions. Then, we refined videos through human-targeted filtering and manual revision. The clips are labeled in two phases, including temporal localization and spatial on-body regions.

3.1 Data Collection and Preprocessing

Video Collection and Clip Preprocessing. Following Bain et al. [3], we constructed HAPACT from movie clips collected from *Movieclips*[1], an official, publicly accessible YouTube channel with stable long-term availability. Movie titles are parsed from video titles and descriptions, and metadata (e.g., release year and genre) are retrieved from IMDb [30]. Non-film works, such as TV shows and documentaries, and adult-rated movies are excluded from data collection.

Genre-based Filtering. Since impact events are not evenly distributed across movie genres, we applied genre-based filtering to prioritize clips more likely to contain salient impact scenes. A genre prior was constructed from films screened in 4DX [12] format, by computing both genre frequencies and genre-pair frequencies. We assigned a genre-prior score to each candidate movie based on its individual genres and genre pairs, then retained movies above the threshold. The detailed scoring formulation and procedure are described in Section A.3.2.

Query-clip Candidate Generation via Vision-Language Model. To generate candidate instances, all movie clips were processed with Gemini 2.5 Pro [13] to identify impact-related moments and produce second-level temporal intervals with textual descriptions of the corresponding events. For each candidate event, a segment was cropped around the estimated interval with a ± 2.5 second buffer, preserving contextual cues before and after the impact. Candidate segments longer than 8 seconds were excluded to focus on short, salient impact events.

Human-targeted Impact Filtering and Manual Revision. Using the generated textual descriptions, human-targeted impacts were identified through two parallel filters. GPT-4o [29] was prompted to identify descriptions in which the impact involves a human, while a Stanford Parser-based filter, inspired by the human-subject filtering of Gupta and Malik [23], retained descriptions whose affected argument denotes a person. The outputs of the two filters were merged to form the candidate pool. The candidates were then manually reviewed and revised by members of the research team to remove cases outside the task scope. To assess whether the candidate pipeline captures human-targeted impacts in the videos, we further evaluated its recall, obtaining a mean recall of 88.99%. Detailed criteria and results are provided in Section A.3.

3.2 Data Annotation

HAPACT was annotated in two phases: temporal impact annotation and spatial on-body impact annotation. We used the HumanSignal [63] platform to conduct both phases with 28 trained annotators, of whom 50% were female. The annotator pool consisted of undergraduate and graduate students across diverse academic disciplines. The process took approximately 30 weeks in total.

Stage 1: Temporal Impact Annotation. In the first phase, annotators labeled the frame-level temporal extent of the queried impact event for each video-query pair. All clips were re-encoded at 24 FPS, and annotators were provided with frame-level access to each clip. The task required marking every occurrence of the queried impact, including repeated instances within the same clip. We

[1] `https://www.youtube.com/@movieclips`

Table 1: **Comparison with Representative Body-Centric Physical Understanding and Video Moment Retrieval Datasets.** #Text Queries, Avg. #M / Query, and Avg. M Length denote the number of queries, average moments per query, and average moment duration, respectively. HAPACT contains shorter moments than existing video grounding datasets, with an average duration of 0.20s.

Dataset	Domain	Scale	#Text Queries	Avg. #M / Query	Avg. M Length	Physical Cue	Body Rep.
Image/video body-level physical datasets							
PROX [24]	Indoor scenes	100,000 frames	✗	✗	✗	Scene contact	SMPL-X
BEHAVE [5]	Natural scenes	15,200 frames	✗	✗	✗	Object contact	SMPL-H
RICH [27]	Natural scenes	577,000 frames	✗	✗	✗	Scene contact	SMPL-X
HOT [11]	In-the-wild	15,082 images	✗	✗	✗	Object/scene contact	2D regions
DAMON [64]	In-the-wild	5,522 images	✗	✗	✗	Object/scene contact	SMPL
3DIR [70]	In-the-wild	5,000 images	✗	✗	✗	Object contact	SMPL-H
BodyContact4D [56]	In-the-wild	800,000 frames	✗	✗	✗	Object/scene contact	24 segments
Video moment retrieval datasets							
TACoS [53]	Cooking	127 videos	18,818	1	6.1s	✗	✗
Charades-STA [16]	Indoor activities	9,848 videos	16,128	1	8.1s	✗	✗
DiDeMo [1]	Open videos	10,464 videos	40,543	1	6.5s	✗	✗
ActivityNet Captions [36]	Open videos	19,811 videos	71,957	1	37.1s	✗	✗
TVR [37]	TV shows	21,800 videos	108,965	1	9.1s	✗	✗
QVHighlights [38]	Vlog / news	10,148 videos	10,310	1.78	24.6s	✗	✗
MAD [57]	Movies	650 movies	384,600	1	4.1s	✗	✗
Ego4D-NLQ [20]	Egocentric	1,659 videos	19,200	1	10.5s	✗	✗
HAPACT	**Movies**	**20,369 annotated frames from 2,884 videos**	**2,884**	**1.47**	**0.2s (4.80 f)**	**Text-conditioned impact events**	**SMPL-X**

collected three independent annotations for each pair and aggregated by majority vote to determine the final temporal intervals. Ambiguous cases after aggregation were resolved through manual adjudication. Section A.4.3 provides the detailed procedure.

Stage 2: Spatial On-body Impact Annotation. In the second stage, annotators labeled the body region affected by the queried impact over time. For each temporally localized impact, annotators viewed the corresponding video segment and marked the affected region on an SMPL-X [51] body mesh for each frame. To support this process, we developed a custom painting interface[2] inspired by prior dense 3D human annotation datasets [64, 70] and deployed it as a web-based tool for fine-grained on-body annotation. The interface allows annotators to directly paint impacted regions on the mesh surface and export consistent vertex-level labels across frames. Additional details on the interface design and annotation workflow are provided in Section A.4.4.

Annotation Quality Assessment. We evaluate annotation quality through inter-annotator agreement and reference-based validation, following Tripathi et al. [64]. These evaluations examine annotator consistency in temporal boundaries and spatial labels, including challenging cases with occlusions and diverse body poses. Detailed procedures and results are provided in Section A.4.5.

3.3 Dataset Analysis

Statistics. To the best of our knowledge, HAPACT is the first human-centric video dataset to provide fine-grained spatio-temporal annotations for physical impact events. Table 1 compares its statistics with existing datasets. Unlike prior datasets, it combines text-conditioned temporal grounding with frame-level on-body impact annotations.

Distribution of Movie Clips and Queries. In total, HAPACT contains $2,884$ video-query pairs with $4,242$ moments, drawn from 212 movies. The dataset is designed around short and salient physical impacts, which is reflected in its temporal statistics. The average moment length is 4.80 frames, corresponding to approximately 0.20 seconds at 24 FPS, and most moments span 1 to 6 frames as shown in Fig. 3(d). A single video-query pair can contain repeated occurrences of the queried impact, resulting in an average of 1.47 moments per query; as shown in Fig. 3(e), multi-moment cases account for 34.1% of the dataset. The relative start and end positions of these impact moments are shown in Fig. 3(b).

[2]A demo of the annotation interface is available at `https://smpl-painter.netlify.app/`.

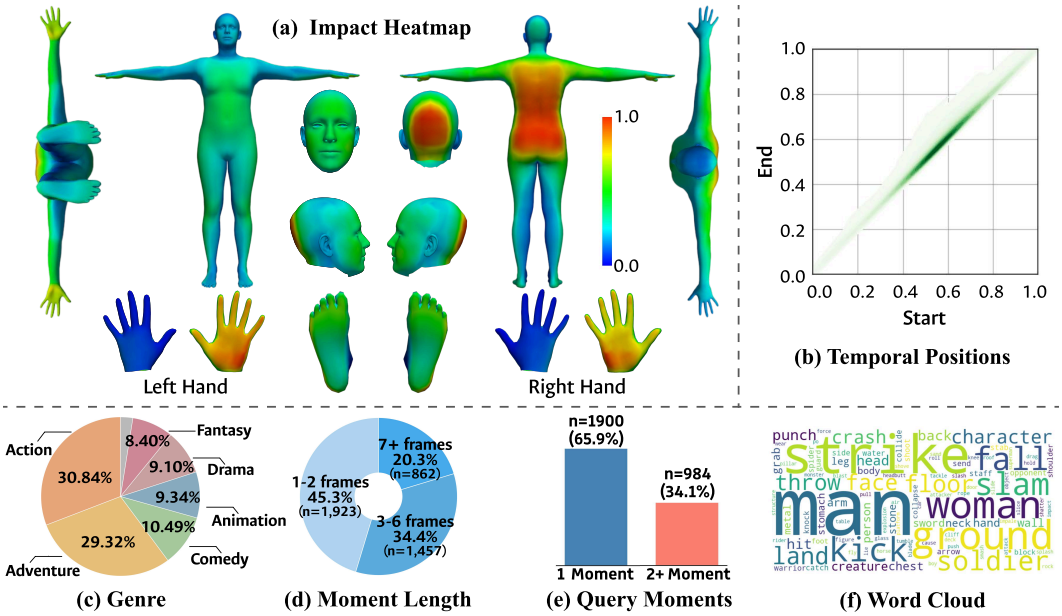


Figure 3: **Data Distribution Overview.** (a) Distribution of impact occurrences on body mesh, where 1 indicates high and 0 indicates low occurrence probability; (b) Distribution of start-end positions of impact moments, normalized from 0 to 1 relative to the total video length; (c) Distribution of movie genres; (d) Distribution of moment lengths by frame count; (e) Distribution of single- and multi-moment queries; (f) Word cloud of verbs and nouns in the description of HAPACT. It shows that verbs such as "strike" and nouns such as "man" appear frequently, emphasizing that the dataset focuses on physical interactions. More detailed statistical information can be found in Section A.5.

The source movie distribution is shaped by the impact-centered construction of HAPACT. As shown in Fig. 3(c), action and adventure genres appear frequently in the collection. Fig. 3(f) shows a word cloud of verbs and nouns appearing in the query sentences, which contain 11.1 words on average.

Distribution of Impact Occurrence on Body Mesh. Fig. 3(a) shows the aggregate vertex-level impact distribution on the SMPL-X mesh across the entire dataset. The head region has the highest intensity (23.3%), followed by the hips (8.4%) and the spine (6.9%). The body-part visualizations in the same panel further show that impacts frequently occur on the back, buttocks, and palms. These patterns reflect common physical impact scenarios in movie clips, such as falls, collisions, and body-ground or body-object contacts.

3.4 Dataset Split

For benchmark evaluation, we partition HAPACT into training, validation, and test splits with a ratio of 70:15:15. The split is movie-disjoint to prevent content leakage, and is constructed to maintain comparable distributions of genre, clip duration, number of moments, moment length, and affected body vertices. The resulting splits contain $2,019$ training pairs with $14,264$ annotated frames, 432 validation pairs with $2,963$ annotated frames, and 433 test pairs with $3,142$ annotated frames. Section A.5.1 reports detailed statistics for each split.

4 Experiments

4.1 Task Definition

Human-centric impact localization is defined over a video with frames $V = \{f_t\}_{t=1}^{T}$ and a textual query q describing a physical impact event involving a human. The objective is to localize both *when* the queried impact occurs and *which body vertices* are affected over time. The task is evaluated along two axes: temporal localization and spatial on-body localization.

Task 1: Temporal Localization. The first task is to predict a set of temporal intervals:

$$\mathcal{T} = \{\tau_k\}_{k=1}^{K}, \quad \tau_k = [t_s^{(k)}, t_e^{(k)}], \quad 1 \le t_s^{(k)} \le t_e^{(k)} \le T,$$

where each interval corresponds to one occurrence of the queried impact event in the video.

Task 2: Spatial On-body Localization. The second task is to predict, for each frame $t \in \bigcup_{k=1}^{K} \tau_k$, a binary on-body label:

$$\mathbf{y}_t \in \{0,1\}^N,$$

where N denotes the number of vertices in the SMPL-X body mesh and $\mathbf{y}_t^{(i)} = 1$ indicates that the i-th vertex is affected by the queried impact at frame t.

4.2 Benchmark Setup

HAPACT provides two benchmark axes based on the task definition: **temporal impact localization** and **spatial on-body impact localization**. In the following sections, we describe the specialized baselines and evaluation metrics for each task, with implementation details included in Section A.6.

4.2.1 Temporal Evaluation

Models. The temporal benchmark includes two types of video moment retrieval (VMR) models: retrieval-based temporal localization models and generation-based models. Retrieval-based models [26, 38, 43] predict temporal intervals by matching video representations with the query, whereas generation-based models [25, 40, 45] directly produce temporal predictions from the video using prompts that incorporate the query text. Overall, we evaluate 18 temporal baselines, comprising 8 retrieval-based models [26, 31, 38, 46, 47, 59, 67, 61] capable of multi-moment prediction and 10 generation-based models spanning VMR-specific models [22, 25, 40, 65, 76], open-source MLLMs [2, 66], and closed-source MLLMs [13, 29, 50].

Metrics. We mainly evaluate temporal localization performance using conventional VMR metrics [38], mean average precision (mAP) and Recall@1 (R1), at a temporal intersection-over-union (tIoU) threshold of 0.3. We use this threshold because impact moments in HAPACT are short, lasting only 0.20 seconds on average, and are sensitive to small boundary shifts. To account for multi-moment queries, we also report the standard metrics adopted in Cao et al. [7]: generalized mean average precision (G-mAP) and mean recall (mR), computed over tIoU thresholds from 0.1 to 0.5 with a step size of 0.1. G-mAP metrics are further reported for single-moment (@1_tgt), two-moment (@2_tgt), and three-or-more-moment (@3+tgt) queries.

4.2.2 Spatial Evaluation

Models. The spatial benchmark includes three human-scene contact (HSC) baselines: BSTRO [27], DECO [64], and GraphiContact [41]. These models estimate body regions in contact from visual input, and are adopted here to examine whether contact representations can transfer to impact localization.

MLLM Baseline. Since the spatial axis lacks established baselines that incorporate text input, we introduce a text-conditioned MLLM baseline that formulates spatial impact localization as query-guided region selection. Inspired by prior studies [11, 56], we convert dense binary labels on SMPL-X into 18 body region groups and provide them as candidate region choices in the prompt. The

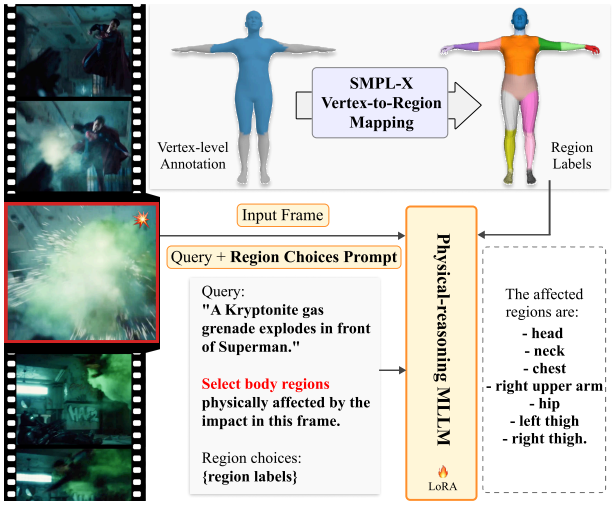


Figure 4: MLLM Baseline for Spatial Impact Localization with Query-Guided Region Selection.

Table 2: **Performance Comparison on Temporal Impact Localization of Retrieval-based Moment Retrieval Models.** *ZS* denotes zero-shot evaluation using QVHighlights [38] checkpoints, and *FT* denotes fine-tuning on the HAPACT training split. *Random* evaluates randomly sampled intervals within each clip. **Bold** indicates the best result and underline indicates the second-best result.

Method	Setting	mAP@0.3	R1@0.3	mAP				mR@k		
				G-mAP	@1_tgt	@2_tgt	@3+tgt	@1	@2	@3
Random		4.14	2.31	7.16	7.30	6.73	5.50	4.11	6.55	7.91
Moment-DETR [38] *NeurIPS '21*	*ZS*	8.05	5.77	10.04	11.69	7.07	5.56	6.93	5.88	7.29
	FT	18.16	13.63	20.39	21.94	19.11	12.92	15.57	16.61	16.59
QD-DETR [47] *CVPR '23*	*ZS*	9.55	6.00	11.98	13.57	9.50	7.81	8.22	8.12	8.53
	FT	21.27	16.17	22.06	23.91	19.98	15.39	16.86	18.67	19.69
EaTR [31] *ICCV '23*	*ZS*	9.49	6.24	12.28	12.85	11.78	6.29	8.13	9.64	11.01
	FT	21.69	16.17	22.54	23.09	22.58	13.92	16.95	19.21	20.62
CG-DETR [46] *arXiv '24*	*ZS*	7.96	4.62	11.09	13.05	8.27	5.08	7.11	6.06	7.29
	FT	20.57	15.47	22.55	23.96	19.48	17.28	17.92	21.52	22.63
SFABD [26] *WACV '24*	*ZS*	5.14	2.08	7.59	8.52	5.82	4.12	3.56	4.55	6.20
	FT	21.19	15.70	20.59	21.91	19.54	12.59	15.79	15.70	17.21
TR-DETR [59] *AAAI '24*	*ZS*	10.62	7.39	14.00	14.97	12.81	6.99	10.21	9.64	11.01
	FT	20.86	16.17	22.20	23.86	20.36	14.62	17.69	17.94	21.39
UVCOM [67] *CVPR '24*	*ZS*	10.20	7.39	13.65	15.56	10.71	5.98	10.67	8.12	8.22
	FT	21.22	**17.32**	23.58	24.28	22.96	16.53	**19.26**	**21.76**	24.65
Sim-DETR [61] *ICCV '25*	*ZS*	5.67	1.85	7.31	7.66	6.75	6.49	3.23	4.79	6.36
	FT	**23.27**	15.47	**26.07**	**26.58**	**24.23**	**25.41**	18.62	**21.76**	**26.82**

Table 3: **Zero-shot Performance Comparison on Temporal Impact Localization of Generation-based Moment Retrieval Models.** Closed-source MLLM results are reported as mean (standard deviation) over five runs. Detailed evaluation settings are provided in Section A.6.2. **Orange** denotes the best result, while purple indicates the second best.

Method	mAP@0.3	R1@0.3	mAP				mR@k		
			G-mAP	@1_tgt	@2_tgt	@3+tgt	@1	@2	@3
VMR-specific Models									
VTimeLLM [25]	1.04	1.15	1.92	2.69	0.82	0.26	2.22	0.00	0.00
TRACE [22]	4.21	4.16	4.76	5.18	3.67	3.25	4.80	6.25	7.50
Grounded-Video-LLM [65]	7.97	10.62	10.06	11.72	8.44	1.37	13.30	0.00	0.00
UniTime [40]	7.93	9.01	12.41	16.42	6.72	1.95	15.06	0.00	0.00
TimeLens [76]	5.10	5.77	9.60	12.54	5.08	2.95	11.83	5.00	0.00
Open Source General MLLMs									
Qwen3-VL-8B [2]	22.88	29.33	23.25	28.31	16.23	7.08	30.07	7.69	11.11
Qwen3-VL-32B [2]	21.45	24.48	22.62	28.90	13.66	5.63	26.50	19.14	6.67
InternVL3.5-8B [66]	14.01	16.17	13.67	14.12	14.63	5.27	15.98	21.18	15.00
Closed Source General MLLMs									
Gemini 3.1 Pro [18]	27.22 (1.34)	29.93 (1.47)	29.70 (1.05)	38.38 (1.89)	16.87 (1.33)	7.21 (1.94)	33.68 (1.28)	28.11 (1.96)	22.56 (4.65)
GPT-4o [29]	28.95 (1.22)	33.67 (1.63)	30.10 (0.76)	37.97 (1.20)	18.30 (0.97)	8.58 (1.64)	34.97 (0.85)	22.62 (2.53)	17.69 (9.63)
GPT-5.4 [50]	**42.64** (0.86)	**47.76** (0.69)	**43.64** (0.74)	**54.49** (1.30)	**26.90** (1.39)	**14.74** (1.88)	**48.80** (0.62)	**38.89** (1.20)	**33.92** (5.02)

model is asked to select all regions physically affected by the given impact query. We instantiate this formulation with Cosmos-Reason2 [49], an MLLM specialized for physical reasoning, and evaluate it after instruction fine-tuning on our training set.

Metrics. Following prior human-scene contact estimation works [27, 41, 64], spatial localization performance is evaluated using frame-wise precision, recall, and F1 score, together with geodesic error in centimeters. For analysis by region, we construct seven body region groups by merging the part segmentation from Chen et al. [11]: head, torso, arms, legs, hips, hands, and feet. For MLLM evaluation, since predictions are made over candidate region labels, we report F1 scores over the predefined labels. Further details are provided in Section A.6.3.

4.3 Main Results

Fine-Tuning on HAPACT improves retrieval-based models on temporal impact localization. For retrieval-based models, Table 2 shows that QVHighlights pre-training does not consistently carry over to temporal impact localization, with several models remaining close to the random baseline.

Table 4: **Performance Comparison on Spatial On-body Impact Localization.** *ZS*, *FT*, and *RS* denote zero-shot, fine-tuning on HAPACT, and fine-tuning with query-guided region selection, respectively. Per-region F1 is reported over seven region groups. **Bold** indicates the best result.

Method	Setting	F1 ↑	Precision ↑	Recall ↑	geo. (cm) ↓	F1 Scores per Body-part ↑						
						Head	Arms	Torso	Hips	Legs	Hands	Feet
BSTRO [27]	*ZS*	6.13	10.69	5.68	81.53	0.00	2.75	0.00	0.35	0.01	14.82	16.61
	FT	5.72	11.20	5.93	51.35	1.65	3.72	7.44	13.40	4.30	7.18	3.66
DECO [64]	*ZS*	8.28	9.87	11.84	83.84	0.00	0.25	1.45	4.21	1.26	12.87	28.86
	FT	5.41	**68.16**	6.18	50.34	1.46	4.78	8.79	10.80	0.85	6.77	0.20
GraphiContact [41]	*ZS*	7.88	8.85	10.23	87.42	0.09	0.28	1.34	8.52	1.84	0.63	29.34
	FT	7.34	30.12	14.18	50.02	5.57	8.48	11.68	15.63	7.19	15.62	6.40
Cosmos-Reason2-2B [49]	*RS*	13.16	11.11	32.41	52.34	10.99	**16.48**	16.13	18.39	10.08	**23.45**	6.80
Cosmos-Reason2-8B [49]	*RS*	**18.59**	16.17	**44.11**	**33.31**	**21.21**	15.92	**20.01**	**29.95**	**12.59**	17.87	**8.71**

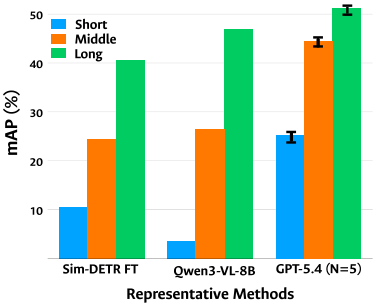


Figure 5: Temporal mAP scores across impact duration groups for representative models.

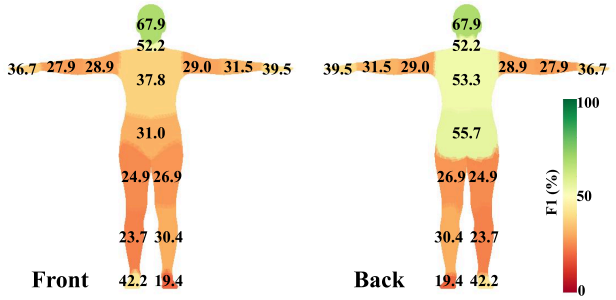


Figure 6: Label-level F1 scores of Cosmos-Reason2-8B fine-tuned using region selection adaptation. Left and right views show the front and back of the SMPL-X body, respectively.

Fine-tuning on HAPACT improves all retrieval models, suggesting that existing VMR architectures can learn to localize impact events with temporal supervision. Still, the best scores remain moderate, reflecting the difficulty of precise localization under short durations and multiple annotated moments.

General MLLMs localize salient impacts but struggle with multi-moments. As presented in Table 3, VMR-targeted generation-based models remain weak on temporal impact localization. In contrast, general MLLMs achieve much stronger zero-shot performance, showing that broad video-language understanding helps identify human impact events. However, even the strongest model (GPT-5.4) drops from mAP@1_tgt to mAP@2_tgt and mAP@3+tgt, decreasing by about 51% and 73%, respectively. This decline indicates that general MLLMs can identify salient impact occurrences, but remain limited in recovering multiple target moments with accurate temporal boundaries.

Temporal performance degrades on short impact moments. Figure 5 further compares length-wise mAP for the strongest representative model from each family. Across all models, short moments (1–2 frames) consistently yield lower mAP than middle (3–6 frames) and long moments (7+ frames). This pattern points to a key difficulty in HAPACT, as current moment retrieval models still have difficulty capturing brief physical impacts. Section A.7.1 presents detailed experiments and analyses of temporal bottlenecks across the evaluated VMR models.

Limited performance of contact-based models on spatial impact localization. Contact-based baselines achieve low accuracy on spatial on-body impact localization, as reported in Table 4. Fine-tuning on HAPACT reduces geodesic errors, but vertex-level F1 scores remain below 9% across all models. These results indicate that physical contact representations alone cannot reliably identify impact-affected body regions from image-only observations. Training from scratch on HAPACT results in similarly low performance, as detailed in Section A.7.2.

Region selection provides a stronger coarse spatial baseline. Cosmos-Reason2-8B, fine-tuned with the region selection formulation, obtains the highest spatial scores in Table 4, despite using coarse body-region predictions rather than vertex-level labels. Still, Figure 6 shows that its label-level

F1 is concentrated on the head, back, and buttocks, while many other regions remain lower. This uneven performance suggests that identifying the affected body region remains challenging even at a coarse level, before considering dense vertex-level localization. Balancing the training-set label distribution narrows but does not eliminate this disparity, as detailed in Section A.7.3.

5 Conclusion

We introduce human-centric impact localization, a new task for identifying queried physical impacts in time and on the human body. We further present HAPACT, the first human-centric video benchmark with fine-grained annotations for physical impact events. Benchmark results show that current models struggle with brief and multi-moment impacts as well as affected-body prediction. Our region selection formulation enables MLLMs for spatial on-body localization, providing a stronger text-conditioned reference while vertex-level estimation remains challenging. HAPACT extends video-based physical reasoning beyond user-intended interactions to externally driven bodily impacts, with potential applications in body-aware video analysis and media-based haptic generation.

Acknowledgments and Disclosure of Funding

The authors declare no competing interests.

References

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

[2] S. Bai, Y. Cai, R. Chen, K. Chen, X. Chen, Z. Cheng, L. Deng, W. Ding, C. Gao, C. Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

[3] M. Bain, A. Nagrani, A. Brown, and A. Zisserman. Condensed movies: Story based retrieval with contextual embeddings. In *Proceedings of the Asian Conference on Computer Vision*, 2020.

[4] S. Bansal, Z. Zhu, S. Tripathi, J. Zhao, M. J. Black, and D. Damen. Towards in-the-wild egocentric 3d hand-object pose estimation. *arXiv preprint arXiv:2606.30598*, 2026.

[5] B. L. Bhatnagar, X. Xie, I. A. Petrov, C. Sminchisescu, C. Theobalt, and G. Pons-Moll. Behave: Dataset and method for tracking human object interactions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15935–15946, 2022.

[6] Z. Cao, H. Du, B. Zhang, X. Yu, X. Li, and S. Wang. When one moment isn't enough: Multi-moment retrieval with cross-moment interactions. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

[7] Z. Cao, H. Du, B. Zhang, X. Yu, X. Li, and S. Wang. When one moment isn't enough: Multi-moment retrieval with cross-moment interactions. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

[8] B. Castellano. PySceneDetect: Python and opencv-based scene cut/transition detection program and library. `https://github.com/Breakthrough/PySceneDetect`, 2014–2026. Accessed: 2026-05-04.

[9] D. Chen and C. D. Manning. A fast and accurate dependency parser using neural networks. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 740–750, 2014.

[10] H. Chen, X. Wang, H. Chen, Z. Zhang, W. Feng, B. Huang, J. Jia, and W. Zhu. Verified: A video corpus moment retrieval benchmark for fine-grained video understanding. *Advances in Neural Information Processing Systems*, 37:40393–40406, 2024.

[11] Y. Chen, S. K. Dwivedi, M. J. Black, and D. Tzionas. Detecting human-object contact in images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17100–17110, 2023.

[12] CJ 4DPLEX. 4DX. `https://www.cj4dplex.com/4dx`, n.d. Accessed: 2026-05-04.

[13] G. Comanici, E. Bieber, M. Schaekermann, I. Pasupat, N. Sachdeva, I. Dhillon, M. Blistein, O. Ram, D. Zhang, E. Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

[14] I. Croitoru, S.-V. Bogolin, S. Albanie, Y. Liu, Z. Wang, S. Yoon, F. Dernoncourt, H. Jin, and T. Bui. Moment detection in long tutorial videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2594–2604, October 2023.

[15] A. Cseke, S. Tripathi, S. K. Dwivedi, A. S. Lakshmipathy, A. Chatterjee, M. J. Black, and D. Tzionas. Pico: Reconstructing 3d people in contact with objects. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 1783–1794, 2025.

[16] J. Gao, C. Sun, Z. Yang, and R. Nevatia. Tall: Temporal activity localization via language query. In *Proceedings of the IEEE international conference on computer vision*, pages 5267–5275, 2017.

[17] G. Gilardi and I. Sharf. Literature survey of contact dynamics modelling. *Mechanism and machine theory*, 37(10):1213–1239, 2002.

[18] Google DeepMind. Gemini 3.1 Pro: Model card. https://deepmind.google/models/model-cards/gemini-3-1-pro/, 2026. Accessed: 2026-05-04.

[19] P. Grady, C. Tang, S. Brahmbhatt, C. D. Twigg, C. Wan, J. Hays, and C. C. Kemp. Pressurevision: estimating hand pressure from a single rgb image. In *European Conference on Computer Vision*, pages 328–345. Springer, 2022.

[20] K. Grauman, A. Westbury, E. Byrne, Z. Chavis, A. Furnari, R. Girdhar, J. Hamburger, H. Jiang, M. Liu, X. Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18995–19012, 2022.

[21] C. Gu, C. Sun, D. A. Ross, C. Vondrick, C. Pantofaru, Y. Li, S. Vijayanarasimhan, G. Toderici, S. Ricco, R. Sukthankar, et al. Ava: A video dataset of spatio-temporally localized atomic visual actions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6047–6056, 2018.

[22] Y. Guo, J. Liu, M. Li, Q. Liu, X. Chen, and X. Tang. Trace: Temporal grounding video llm via causal event modeling. *arXiv preprint arXiv:2410.05643*, 2024.

[23] S. Gupta and J. Malik. Visual semantic role labeling. *arXiv preprint arXiv:1505.04474*, 2015.

[24] M. Hassan, V. Choutas, D. Tzionas, and M. J. Black. Resolving 3d human pose ambiguities with 3d scene constraints. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2282–2292, 2019.

[25] B. Huang, X. Wang, H. Chen, Z. Song, and W. Zhu. Vtimellm: Empower llm to grasp video moments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14271–14280, 2024.

[26] C. Huang, Y.-L. Wu, H.-H. Shuai, and C.-C. Huang. Semantic fusion augmentation and semantic boundary detection: A novel approach to multi-target video moment retrieval. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 6783–6792, 2024.

[27] C.-H. P. Huang, H. Yi, M. Höschle, M. Safroshkin, T. Alexiadis, S. Polikovsky, D. Scharstein, and M. J. Black. Capturing and inferring dense full-body human-scene contact. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13274–13285, 2022.

[28] Y. Huang, O. Taheri, M. J. Black, and D. Tzionas. Intercap: Joint markerless 3d tracking of humans and objects in interaction. In *DAGM German Conference on Pattern Recognition*, pages 281–299. Springer, 2022.

[29] A. Hurst, A. Lerer, A. P. Goucher, A. Perelman, A. Ramesh, A. Clark, A. Ostrow, A. Welihinda, A. Hayes, A. Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

[30] IMDb. IMDb non-commercial datasets. https://developer.imdb.com/non-commercial-datasets/, 2026. Accessed: 2026-05-04.

[31] J. Jang, J. Park, J. Kim, H. Kwon, and K. Sohn. Knowing where to focus: Event-aware transformer for video grounding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13846–13856, 2023.

[32] N. Jiang, T. Liu, Z. Cao, J. Cui, Z. Zhang, Y. Chen, H. Wang, Y. Zhu, and S. Huang. Full-body articulated human-object interaction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9365–9376, 2023.

[33] K. John and H. Seifi. Hapticlens: Interactive vibrotactile haptic generation from spatially localized video motion. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*, pages 1–21, 2026.

[34] K. Jung, S. Kim, S. Oh, and S. H. Yoon. Hapmotion: motion-to-tactile framework with wearable haptic devices for immersive vr performance experience. *Virtual Reality*, 28(1):13, 2024.

[35] R. Khirodkar, J.-T. Song, J. Cao, Z. Luo, and K. Kitani. Harmony4d: A video dataset for in-the-wild close human interactions. *Advances in Neural Information Processing Systems*, 37: 107270–107285, 2024.

[36] R. Krishna, K. Hata, F. Ren, L. Fei-Fei, and J. Carlos Niebles. Dense-captioning events in videos. In *Proceedings of the IEEE international conference on computer vision*, pages 706–715, 2017.

[37] J. Lei, L. Yu, T. L. Berg, and M. Bansal. Tvr: A large-scale dataset for video-subtitle moment retrieval. In *European Conference on Computer Vision*, pages 447–463. Springer, 2020.

[38] J. Lei, T. L. Berg, and M. Bansal. Detecting moments and highlights in videos via natural language queries. *Advances in Neural Information Processing Systems*, 34:11846–11858, 2021.

[39] Y. Li, Y. Yoo, A. Weill-Duflos, and J. Cooperstock. Towards context-aware automatic haptic effect generation for home theatre environments. In *Proceedings of the 27th ACM symposium on virtual reality software and technology*, pages 1–11, 2021.

[40] Z. Li, S. Di, Z. Zhai, W. Huang, Y. Wang, and W. Xie. Universal video temporal grounding with generative multi-modal large language models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

[41] X. Lin, Y. Shen, J. Ma, Y. Sun, C. Bu, W. Zhang, Z. Zhang, H. Fei, L. Jin, and H. Zhao. Graphicontact: Pose-aware human-scene robust contact perception for interactive systems. *arXiv preprint arXiv:2603.20310*, 2026.

[42] Y. Liu, L. Wang, Y. Wang, X. Ma, and Y. Qiao. Fineaction: A fine-grained video dataset for temporal action localization. *IEEE transactions on image processing*, 31:6937–6950, 2022.

[43] Y. Liu, J. He, W. Li, J. Kim, D. Wei, H. Pfister, and C. W. Chen. R 2-tuning: Efficient image-to-video transfer learning for video temporal grounding. In *European conference on computer vision*, pages 421–438. Springer, 2024.

[44] J. Lu, C.-H. P. Huang, U. Bhattacharya, Q. Huang, and Y. Zhou. Humoto: A 4d dataset of mocap human object interactions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10886–10897, 2025.

[45] B. Meinardus, H. G. Rodriguez, A. Batra, A. Rohrbach, and M. Rohrbach. Chrono: A simple blueprint for representing time in mllms. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4092–4097, 2025.

[46] W. Moon, S. Hyun, S. Lee, and J.-P. Heo. Correlation-guided query-dependency calibration for video temporal grounding. *arXiv preprint arXiv:2311.08835*, 2023.

[47] W. Moon, S. Hyun, S. Park, D. Park, and J.-P. Heo. Query-dependent video representation for moment retrieval and highlight detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 23023–23033, 2023.

[48] T. Nishimura, S. Nakada, H. Munakata, and T. Komatsu. Lighthouse: A user-friendly library for reproducible video moment retrieval and highlight detection. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 53–60, 2024.

[49] NVIDIA. Cosmos-Reason2: A multimodal model for physical and commonsense reasoning. Technical report, NVIDIA, 2025. URL `https://huggingface.co/nvidia/Cosmos-Reason2-8B`. Model card. Accessed: 2026-05-04.

[50] OpenAI. GPT-5.4 Thinking System Card. `https://openai.com/index/gpt-5-4-thinking-system-card/`, 2026. Accessed: 2026-05-04.

[51] G. Pavlakos, V. Choutas, N. Ghorbani, T. Bolkart, A. A. Osman, D. Tzionas, and M. J. Black. Expressive body capture: 3d hands, face, and body from a single image. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10975–10985, 2019.

[52] T. Perrett, A. Darkhalil, S. Sinha, O. Emara, S. Pollard, K. Parida, K. Liu, P. Gatti, S. Bansal, K. Flanagan, J. Chalk, Z. Zhu, R. Guerrier, F. Abdelazim, B. Zhu, D. Moltisanti, M. Wray, H. Doughty, and D. Damen. Hd-epic: A highly-detailed egocentric video dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2025.

[53] M. Regneri, M. Rohrbach, D. Wetzel, S. Thater, B. Schiele, and M. Pinkal. Grounding action descriptions in videos. *Transactions of the Association for Computational Linguistics*, 1:25–36, 2013.

[54] A. Sadhu, T. Gupta, M. Yatskar, R. Nevatia, and A. Kembhavi. Visual semantic role labeling for video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5589–5600, 2021.

[55] K. J. Seo, J. Seo, H. Jeong, S. Kim, and S. H. Yoon. Posture-informed muscular force learning for robust hand pressure estimation. In *NeurIPS*, 2024.

[56] S. Shin, C. Lee, H. Chen, J.-T. Song, E. Halilaj, and K. Kitani. Bodycontact4d: A multi-view video dataset for understanding human and environment interactions. In *Thirteenth International Conference on 3D Vision*, 2026.

[57] M. Soldan, A. Pardo, J. L. Alcázar, F. Caba, C. Zhao, S. Giancola, and B. Ghanem. Mad: A scalable dataset for language grounding in videos from movie audio descriptions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5026–5035, 2022.

[58] E. Song, W. Chai, G. Wang, Y. Zhang, H. Zhou, F. Wu, H. Chi, X. Guo, T. Ye, Y. Zhang, et al. Moviechat: From dense token to sparse memory for long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18221–18232, 2024.

[59] H. Sun, M. Zhou, W. Chen, and W. Xie. Tr-detr: Task-reciprocal transformer for joint moment retrieval and highlight detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 4998–5007, 2024.

[60] O. Taheri, N. Ghorbani, M. J. Black, and D. Tzionas. Grab: A dataset of whole-body human grasping of objects. In *European conference on computer vision*, pages 581–600. Springer, 2020.

[61] J. Tang, Z. Wei, Y. Zhu, C. Shi, G. Li, L. Lin, and S. Yang. Sim-detr: Unlock detr for temporal sentence grounding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22760–22771, 2025.

[62] Z. Tang, Y. Liao, S. Liu, G. Li, X. Jin, H. Jiang, Q. Yu, and D. Xu. Human-centric spatio-temporal video grounding with visual transformers. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(12):8238–8249, 2022.

[63] M. Tkachenko, M. Malyuk, A. Holmanyuk, and N. Liubimov. Label Studio: Data labeling software. `https://github.com/HumanSignal/label-studio`, 2020–2025. Open source software available from `https://github.com/HumanSignal/label-studio`. Accessed: 2026-05-04.

[64] S. Tripathi, A. Chatterjee, J.-C. Passy, H. Yi, D. Tzionas, and M. J. Black. Deco: Dense estimation of 3d human-scene contact in the wild. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8001–8013, 2023.

[65] H. Wang, Z. Xu, Y. Cheng, S. Diao, Y. Zhou, Y. Cao, Q. Wang, W. Ge, and L. Huang. Grounded-VideoLLM: Sharpening fine-grained temporal grounding in video large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 959–975, 2025.

[66] W. Wang, Z. Gao, L. Gu, H. Pu, L. Cui, X. Wei, Z. Liu, L. Jing, S. Ye, J. Shao, et al. Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025.

[67] Y. Xiao, Z. Luo, Y. Liu, Y. Ma, H. Bian, Y. Ji, Y. Yang, and X. Li. Bridging the gap: A unified video comprehension framework for moment retrieval and highlight detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18709–18719, 2024.

[68] X. Xie, B. L. Bhatnagar, and G. Pons-Moll. Chore: Contact, human and object reconstruction from a single rgb image. In *European Conference on Computer Vision*, pages 125–145. Springer, 2022.

[69] L. Xu, X. Lv, Y. Yan, X. Jin, S. Wu, C. Xu, Y. Liu, Y. Zhou, F. Rao, X. Sheng, et al. Inter-x: Towards versatile human-human interaction analysis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 22260–22271, 2024.

[70] Y. Yang, W. Zhai, H. Luo, Y. Cao, and Z.-J. Zha. Lemon: Learning 3d human-object interaction relation from 2d images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16284–16295, 2024.

[71] Y. Yang, W. Zhai, C. Wang, C. Yu, Y. Cao, and Z.-J. Zha. Egochoir: Capturing 3d human-object interaction regions from egocentric views. *Advances in Neural Information Processing Systems*, 37:54529–54557, 2024.

[72] Y. Yin, C. Guo, M. Kaufmann, J. J. Zarate, J. Song, and O. Hilliges. Hi4d: 4d instance segmentation of close human interaction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17016–17027, 2023.

[73] Q. You, Q. Wu, Y. Li, W. Ji, L. Li, P. Cai, L. Wei, and R. Zimmermann. Generalized video moment retrieval. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu, editors, *International Conference on Learning Representations*, volume 2025, pages 49207–49223, 2025.

[74] H. Yuan, J. Ni, Y. Wang, J. Zhou, Z. Liang, Z. Liu, Z. Cao, Z. Dou, and J.-R. Wen. Momentseeker: A comprehensive benchmark and a strong baseline for moment retrieval within long videos. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

[75] A. Zala, J. Cho, S. Kottur, X. Chen, B. Oğuz, Y. Mehdad, and M. Bansal. Hierarchical video-moment retrieval and step-captioning. In *CVPR*, 2023.

[76] J. Zhang, T. Wang, Y. Ge, Y. Ge, X. Li, Y. Shan, and L. Wang. Timelens: Rethinking video temporal grounding with multimodal llms. *arXiv preprint arXiv:2512.14698*, 2025.

[77] Z. Zhang, Z. Zhao, Y. Zhao, Q. Wang, H. Liu, and L. Gao. Where does it exist: Spatio-temporal video grounding for multi-form sentences. In *CVPR*, 2020.

[78] Y. Zhao, T. Kwon, P. Streli, M. Pollefeys, and C. Holz. Egopressure: A dataset for hand pressure and pose estimation in egocentric vision. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 27727–27738, 2025.

A Appendix

A.1 Broader Impacts

HAPACT broadens video-based physical understanding by treating impact as a human-centric physical cue. While many interaction datasets focus on intentional human actions, impact events often involve external forces, object collisions, and environmental contact. This perspective can support research on how dynamic events affect the human body in visual media, with possible applications in physical event analysis, body-aware video understanding, haptic generation for audiovisual content, and mixed reality experiences. The annotation pipeline may also be adapted to other video domains that require human-centered physical event annotations.

The potential negative impact is limited because HAPACT is built from staged movie clips and does not redistribute the underlying copyrighted footage. However, the dataset contains scenes with violence, falls, collisions, and explosions. Models trained or evaluated on this benchmark should therefore be used with care when transferred beyond the domain, where predictions may be misinterpreted without appropriate context.

A.2 Video Availability

HAPACT is constructed from videos collected from the *Movieclips* YouTube channel, following prior movie-based datasets that use the same source [3, 54]. This source offers relatively stable availability compared to YouTube videos uploaded by individual users. For example, Bain et al. [3] reported that only 0.3% of videos were removed from the *Movieclips* channel between January 1 and September 1, 2020. To support reproducibility, we provide source video identifiers, trimmed timestamps, IMDb identifiers for metadata, annotations, and evaluation scripts, while the underlying movie footage remains owned by the original rights holders.

A.3 Dataset Construction Details

A.3.1 Video Collection and Preprocessing

We constructed our initial video pool from the *Movieclips* YouTube channel, using videos available as of August 12, 2025. To remove promotional outros while avoiding unnecessary trimming, we applied the fixed-duration trimming rule of Bain et al. [3] only when an outro was detected. Specifically, we segmented each video into shots using PySceneDetect [8] and compared the terminal shot with reference outro clips based on visual similarity. If an outro match was found, we removed the final 10 seconds for videos uploaded before May 2017 and the final 30 seconds for videos uploaded from May 2017 onward; otherwise, the video was left untrimmed.

A.3.2 Genre-Based Scoring Prior from 4DX Films

To construct a video pool rich in physical impact events, we selected candidate movies whose genres were similar to those of films previously screened in 4DX. We compiled a list of 4DX-screened films as of August 7, 2025, and computed how frequently each genre and genre pair appeared among them based on IMDb genre metadata. For each candidate movie m, let G_m denote its set of IMDb genres assigned to m. The single-genre score is defined as

$$S_{\text{single}}(m) = \frac{1}{|G_m|} \sum_{g \in G_m} p(g),$$

where $p(g)$ is the empirical probability of genre g among the 4DX-screened films. For movies with at least two valid genres, we define the genre-pair score as

$$S_{\text{pair}}(m) = \frac{1}{|\mathcal{P}_m|} \sum_{(g_i, g_j) \in \mathcal{P}_m} p(g_i, g_j),$$

where $\mathcal{P}_m$ denotes the set of genre pairs in G_m, and $p(g_i, g_j)$ is the empirical probability of the genre pair among the 4DX-screened films. The final genre-based score is defined as

$$S(m) = S_{\text{single}}(m) + S_{\text{pair}}(m).$$

Movies with no valid genres were assigned a score of zero, and $S_{\text{pair}}(m)$ was set to zero when fewer than two valid genres were available. Candidate movies satisfying $S(m) \geq 0.1$ were retained.

A.3.3 Query-Clip Candidate Generation

After filtering the candidate movie pool, we generated candidate query-clip pairs for physical impact events with Gemini 2.5 Pro [13]. Given each video, the model was prompted to identify second-level temporal segments in which a person experiences a physical impact, and to return each segment with a short text description. The prompt was as follows:

Prompt for Query-Clip Candidate Generation

Identify segments where a person experiences a physical impact. Consider only sudden and salient physical forces affecting the subject. Exclude light contact, minor bumps, distant events, off-screen events, or events that are not physically relevant to the affected subject.

Answer in the format "[MM:SS - MM:SS] Description of the impact". If there are multiple events, return multiple lines. If none, answer "None".

The returned descriptions formed the initial impact queries. Since the model outputs second-level timestamps in the `MM:SS` format, each segment was expanded with a 2.5-second temporal buffer on both sides. If the buffer exceeded the clip boundaries, the interval was clipped to remain within the original video and aligned to a 0.5-second step. The resulting intervals were extracted and re-encoded at 24 FPS to form the query-clip candidate pairs.

A.3.4 Human-Targeted Impact Filtering

To remove impact descriptions that do not involve humans, we applied two complementary filters. First, GPT-4o [29] was prompted to determine whether each impact physically involves a human, returning a true-or-false label. The prompt counted a description as human-involved when a human was the recipient or direct cause of the impact, or was physically affected by the resulting force:

Prompt for Human-Targeted Impact Filtering

Determine whether the described impact physically involves a human. Count as True if a human is the recipient of the impact, causes the impact through direct physical action, or is directly affected by the resulting physical force. Do not count cases where a human only observes, hears, or is merely near the impact without being physically affected.

Answer with only "True" or "False".

Second, we applied a Stanford Parser-based filter [9] to identify descriptions whose affected argument denotes a person. Following the human-subject filtering of Gupta and Malik [23], we parsed each description and checked whether the argument associated with the impact predicate matched a list of person-denoting nouns and pronouns. We retained the union of query-clip pairs accepted by the two filters.

A.3.5 Recall Evaluation of the Candidate Pipeline

To assess the recall of the candidate generation and filtering steps, we conducted an additional evaluation with authors blinded to the candidate outputs. Pairs of authors annotated all human-targeted impacts across the *Movieclips* videos used in HAPACT construction at one-second resolution and cross-checked their annotations. We then compared these annotations with the candidate results retained after human-targeted impact filtering. The resulting mean recall was 88.99% (95% confidence interval: 87.98–90.00%), indicating that the pipeline captures most human-targeted impacts present in the source videos.

A.3.6 Manual Revision

Following human-targeted filtering, author pairs reviewed all query-clip candidates and excluded cases outside the task scope:

- **Impact event affecting a single target person.** As an exception, two-person impacts were retained only when the event involved direct human-human impact.

- **Impact event visible and reliably inferable.** Ambiguous, extremely off-screen, severely blurred, or occluded impacts were excluded.
- **Physically plausible impact event.** Non-physical or supernatural impacts were discarded.

A candidate was accepted only when both authors judged it to satisfy these criteria. When a valid candidate instance could be obtained by editing the query, one author proposed the revision and the other reviewed it for approval. If both authors proposed different revisions for the same instance, the final query was selected by consensus.

A.4 Dataset Annotation Details

A.4.1 Annotator Selection and Distribution

Annotators were selected separately for each annotation task after a training phase. For each task, applicants annotated a screening set and their results were evaluated against author-annotated pseudo ground truth. Among 52 applicants, 28 annotators were selected. The selected annotators included 14 female annotators, with 7 graduate students and 21 undergraduate students. Annotators were compensated at a rate of $\$0.12$ per video-query pair for temporal annotation and $\$1.50$ per video-query pair for spatial annotation.

The annotation protocol was approved by the KAIST Institutional Review Board (IRB No. KAISTIRB-2025-235). Annotators were informed before participation that the task involved viewing movie clips with physical impact events, including violence, falls, collisions, and explosions, which could cause discomfort or fatigue.

A.4.2 Annotation Instructions

Annotators were provided with task-specific instruction materials before each annotation stage. They were trained on a definition of physical impact based on Gilardi and Sharf [17]. Following this definition, annotators identified each occurrence of the queried impact, marking its temporal extent and the body region affected in each impact frame. The core instructions from the materials are reproduced below.

Instructions for Temporal Impact Annotation

You are given a movie clip and an impact description.

Your task is to mark the start and end frames of the physical impact described in the video. Review the impact description and identify when the impact is applied to the person. Mark only the impact timing itself, excluding the approach before the impact and the aftermath after the impact. If the described impact occurs multiple times, mark each occurrence separately.

If the queried impact is not visible or cannot be reliably inferred, leave the comment and skip.

Instructions for Spatial On-body Impact Annotation

You are given a movie clip, an impact description, and the temporal interval(s) for the impact.

Carefully review the clip, the impact description, and the provided interval(s). For every frame within the interval(s), identify which part of the human body is affected by the impact and mark the affected region using the web-based annotation tool. All frames inside the provided interval(s) must be annotated. For human-human impacts, use the grammatical subject of the impact description as the target person, and annotate only the body region of that subject.

If the impact is not perceived on the subject in a frame, or if the affected body region cannot be inferred, briefly explain the reason in the comment and remove that frame from the interval.

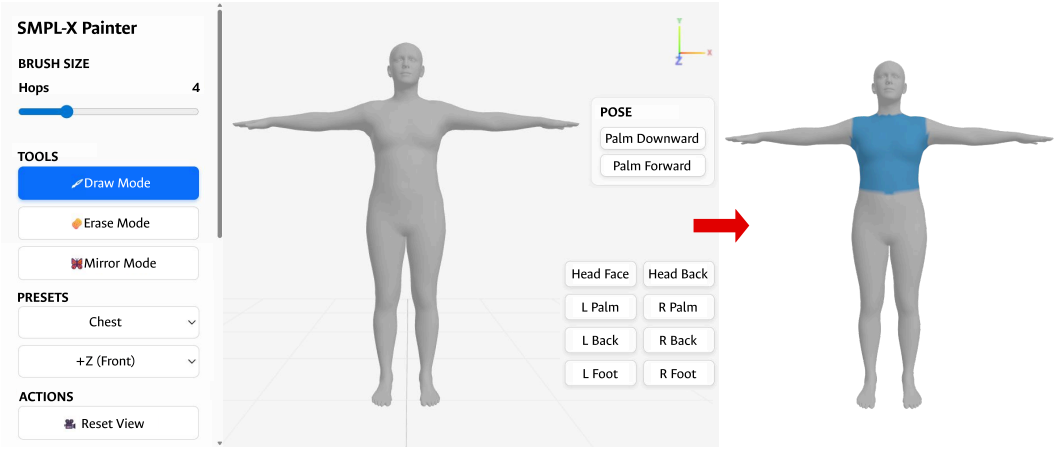


Figure 7: **Spatial annotation interface.** Annotators mark affected body regions on an SMPL-X mesh using draw and erase modes, with controls for mesh navigation, brush size, and shortcut views.

A.4.3 Temporal Annotation Aggregation

For temporal annotation, each video-query pair was annotated by three annotators. To aggregate them, each annotated interval was first converted into a frame-level binary label. The final temporal intervals were obtained by majority voting, where a frame was retained if at least two of the three annotators labeled it as containing the queried event. Ambiguous cases after aggregation were resolved by author consensus. These included cases with low agreement among annotators, measured by a mean pairwise tIoU of at most 0.5, and cases where majority voting collapsed the annotated intervals into fewer moments.

A.4.4 Spatial Annotation Interface

For spatial annotation, we built a dense on-body 3D annotation interface inspired by prior 3D human contact annotation datasets [64, 70]. The interface centers on an SMPL-X mesh that can be translated, rotated, and zoomed. The left panel provides controls for switching between draw and erase modes, as well as a slider for adjusting the brush size. In draw mode, annotators mark affected regions by clicking and dragging on the mesh, while in erase mode, they remove marked regions. The lower-right panel provides shortcut buttons for specific views, including the palms, feet, face, and back of the head. To reduce ambiguity caused by differences between the canonical and observed poses, the annotation interface provided multiple canonical T-pose variants. The interface was deployed as a web-based tool to be used alongside the HumanSignal platform. Annotators reviewed the video, query, and temporal intervals for each assigned data instance in HumanSignal, marked the affected body regions for each frame using the interface, and copied the resulting vertex indices back to HumanSignal. The interface design is shown in Fig. 7. Impact intervals for which annotators could not infer the affected body region with sufficient confidence were excluded from the final dataset.

A.4.5 Quality Control and Assessment

To maintain annotation quality, we applied two quality-control procedures across both annotation stages. First, as described in Section A.4.1, annotators were selected through a pre-screening task in which their annotations were evaluated against author-annotated references. Second, all submitted annotations were reviewed by meta-annotators, and low-quality cases were returned for re-annotation with explanatory feedback.

Furthermore, annotation quality was evaluated using both inter-annotator agreement and reference-based validation, following the protocol of Tripathi et al. [64]. For temporal annotation, annotators labeled a shared set of 10 impact moments from 8 movies to measure agreement, yielding a Fleiss' κ of 0.46. Fleiss' κ ranges from -1 to 1, with -1, 0, and 1 indicating complete disagreement, chance-level agreement, and complete agreement, respectively. A value of 0.46 indicates moderate agreement. To further validate boundary accuracy on existing data, a curated set of 100 physical-

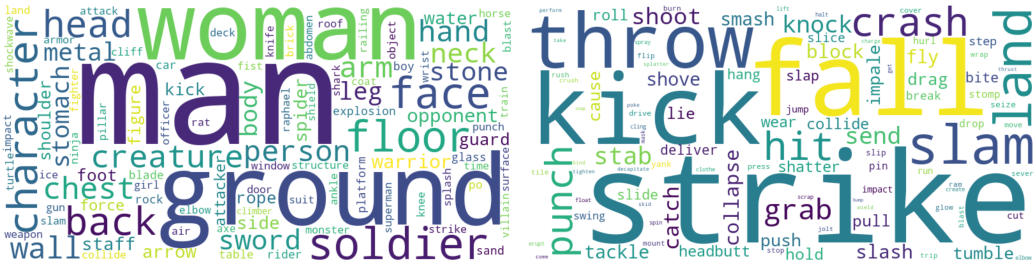


Figure 8: Distribution of nouns (left) and verbs (right) in the descriptions.

	Total	Train	Test	Val
Video Clip Statistics				
Number of Movies	212	148	32	32
Number of Clips	2,884	2,019	433	432
Mean Clip Duration	6.49s	6.49s	6.44s	6.55s
Query Statistics				
Total Word Count	32,039	22,395	4,920	4,724
Mean Word Count per Query	11.1	11.1	11.4	10.9
Moment Length Statistics				
Number of Moments	4,242	2,956	649	637
Mean Frames Count per Moment	4.80	4.83	4.84	4.65
Short (1–2 frames)	45.33%	45.53%	44.68%	45.05%
Middle (3–6 frames)	34.35%	34.10%	34.51%	35.32%
Long (7+ frames)	20.32%	20.37%	20.80%	19.62%

Table 5: Statistics on splits of HAPACT.

interaction instances from FineAction [42] was annotated under the same protocol. The resulting mean temporal IoU was 0.70, indicating reliable temporal boundary annotation.

For spatial annotation, annotators labeled 20 shared frames from 7 movies to measure agreement. Fleiss' κ was computed over all SMPL-X vertices, yielding 0.53 at the vertex level. We additionally measured body-part-level agreement by aggregating vertices according to the official SMPL-X body segmentation provided by Meshcapade[3], which yielded a Fleiss' κ of 0.61. These results show stronger consistency at the body-part level, with the vertex-level score comparable to those reported in representative human-annotated physical-interaction datasets [4, 64].

To assess whether annotators could consistently produce labels from images with occlusions and diverse body poses, we evaluated their annotations against sensor-derived 3D references on 100 samples. We followed prior work [64], which also uses canonical-mesh-based annotations. The data were curated from PROX [24] and RICH [27], and annotated following the human-scene contact task protocol to evaluate consistency with external mesh-based reference labels. For these samples, we manually created text queries to constrain which body region should be marked. The resulting mIoU with the reference labels was 0.42, comparable to the value reported by Tripathi et al. [64].

A.5 More Dataset Statistics

A.5.1 Overall Dataset Statistics

Figure 8 shows the distribution of frequent nouns and verbs extracted from query descriptions. Frequent nouns such as "man" and "woman" reflect the human-centric nature of the dataset. Verbs such as "strike" and "kick" indicate that the dataset primarily deals with direct physical impact.

[3]Available at `https://github.com/Meshcapade/wiki/tree/main/assets/SMPL_body_segmentation/smplx`

Moments per Query	Total	Train	Test	Val
1	1,900	1,355	272	273
2	753	507	119	127
3+	231	157	42	32
Total	2,884	2,019	433	432

Table 6: Distribution of the number of moments per query for each split.

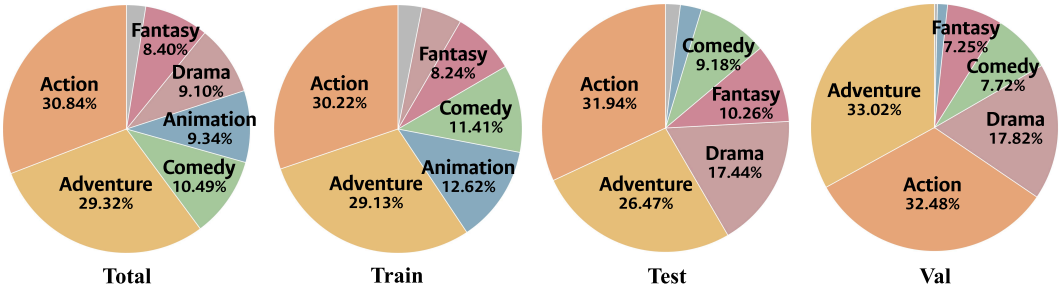


Figure 9: **Distribution of movie genres for each split.** This shows that action and adventure comprise a similar proportion of each split.

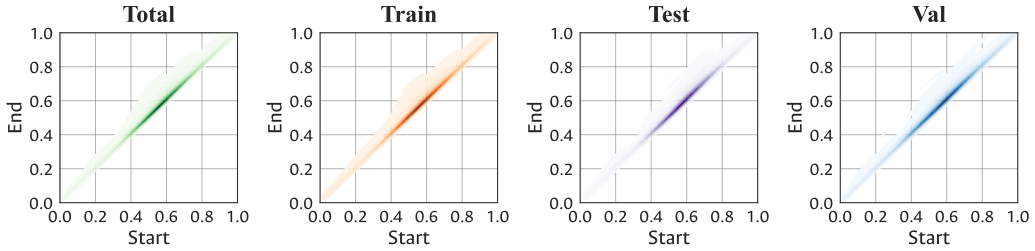


Figure 10: **Distribution of start-end positions for each split.** Start and end positions of impact moments, normalized from 0 to 1 relative to the total video length. The distribution of moment positions is comparable across all splits.

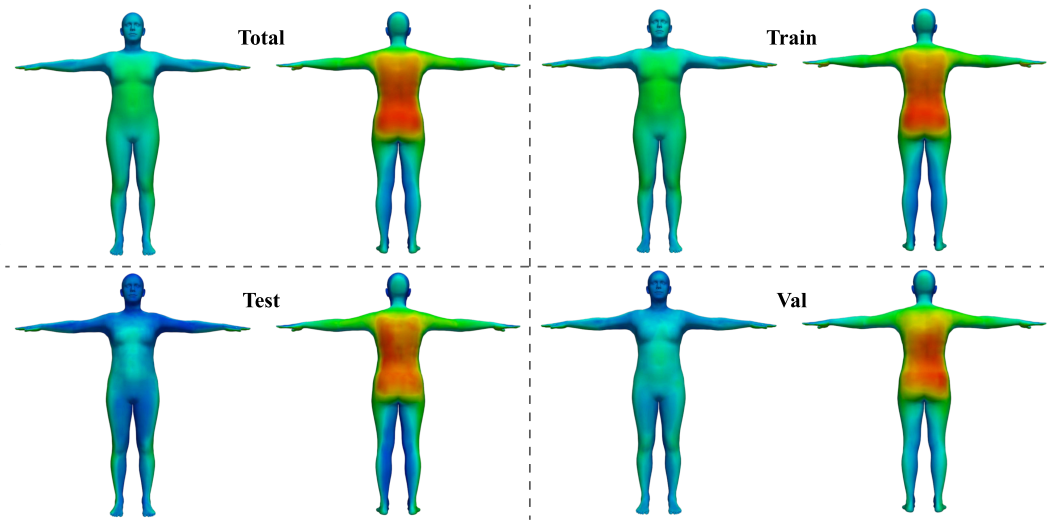


Figure 11: **Distribution of impact occurrences on the body mesh for each split.** 1 indicates a high probability of contact and 0 indicates a low probability. This indicates that the affected body vertices were distributed similarly across the splits.

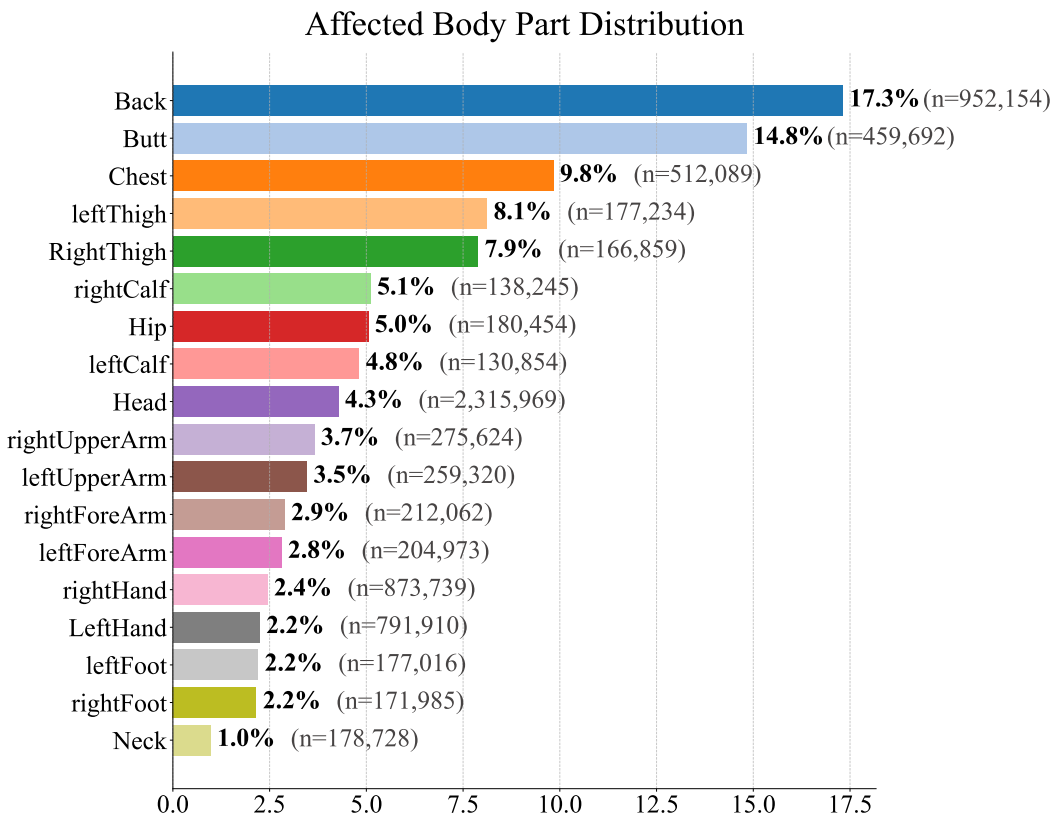


Figure 12: **Distribution of affected body parts weighted by face-level intensity across the SMPL-X mesh.** Each bar represents the face-level intensity score (%) for each body part.

A.5.2 Split Statistics

Table 5 shows the statistics for each split, across the clip, query, and moment distributions. This confirms that the splits are structured to maintain consistent distributions of clip duration, number of moments, and moment length. Table 6 shows the distribution of the number of moments per query. The dataset includes queries with an average of 1.47 moments, as well as a significant proportion of queries with two or more moments.

A.6 Evaluation Details

A.6.1 General Evaluation Settings

Evaluation was conducted on a server equipped with 4 NVIDIA RTX A6000 GPUs, each with 48GB of memory. Benchmark models were evaluated on the HAPACT test split, while models requiring training or fine-tuning were trained on the training split and selected using the validation split.

A.6.2 Temporal Localization Evaluation

For retrieval-based baselines, CLIP features were extracted at a sampling interval of 0.05 seconds to provide dense frame coverage for each clip. Both visual and query features were obtained following the CLIP feature extraction pipeline provided by Lighthouse [48]. Evaluation was conducted using the same framework and configurations, which provide a reproducible codebase for video moment retrieval. Since Sim-DETR and SFABD are not included in Lighthouse, they were evaluated using their official repositories. Models were initialized from weights pretrained on QVHighlights and fine-tuned for 200 epochs with a learning rate of 1×10^{-5}. For methods using saliency supervision, the saliency loss weight was set to 1.0.

Table 7: **Prompts for Generation-based Temporal Localization Baselines.** Each prompt follows the model's native input convention while preserving the same impact-localization constraints.

Model	Prompt
VTimeLLM	This is a video with {num_frames} frames: `<video>` During which frames can we see the impact event "{query}" happening in the video? If the event occurs multiple separate times, return all intervals, not just the most salient one. Localize only the applied moment of impact, not the approach or aftermath. Return all intervals in chronological order. Answer only in the format: From s1 to e1; From s2 to e2.
TRACE	`<video>` Localize a series of moments in the video where the impact on a person's body described by "{query}" is physically applied, and output the start and end timestamps for each matching moment. The query may match multiple separate impact moments in the same video. Return all matching intervals, not just the most salient one. Localize only the applied moment of impact, not the approach or aftermath.
Grounded-Video-LLM	`<video> <grounding>` Give you a textual query: "{query}". The described content may occur multiple times in the video, where a physical impact is applied to a person's body. When do all the described impact moments occur in the video? Localize only the moment when the impact is physically applied to the person, not the approach or aftermath. Please return all start and end timestamps in chronological order.
UniTime	This is a sequence interleaved with timestamps and frames. Your task is to identify the temporal window(s) (start and end timestamps) when the physical impact on a person's body described by the given query is applied. The query may match multiple separate impact moments in the video. Localize only the applied moment of impact, not the approach or aftermath. Query: {query} Answer:
TimeLens	You are given a video with multiple frames. The numbers before each video frame indicate its sampling timestamp (in seconds). Please find the physical impact on a person's body described by the sentence "{query}", determining its starting and ending times. The query may match multiple separate impact moments in the same video. Localize only the moment of impact, not the approach or aftermath. The format should be: "The event happens in <start time 1> - <end time 1> seconds; <start time 2> - <end time 2> seconds; ...".
Qwen3-VL / InternVL-3.5	Identify one or more time intervals in this video where the impact on a person's body described by "{query}" is physically applied. The query may match multiple separate impact moments in the same video. Return all matching intervals, not just the most salient one. Localize only the moment when the impact is applied to the person, not the approach or aftermath. Return only the list of relevant intervals in seconds as `[[start_1, end_1], ..., [start_n, end_n]]`, for example `[[1.040, 1.375], [4.833, 5.125], [7.208, 7.417]]`. If there is only one interval, return it as a nested list `[[start, end]]`. Do not return any explanation or extra text.
Gemini 3.1 Pro	Localize one or more time intervals in this video where the impact on a person's body described by "{query}" is physically applied. The query may match multiple separate impact moments in the same video. Return all matching intervals, not just the most salient one. Localize only the moment when the impact is applied to the person, not the approach or aftermath. Return only the list of relevant intervals in the format `[[MM:SS.SSS, MM:SS.SSS], ..., [MM:SS.SSS, MM:SS.SSS]]`. For example `[[00:01.040, 00:01.375], [00:04.833, 00:05.125], [00:07.208, 00:07.417]]`. If there is only one interval, return it as a nested list `[[MM:SS.SSS, MM:SS.SSS]]`. Do not return any explanation or extra text.
GPT-4o / GPT-5.4	You are given multiple frames from a video. Each frame is preceded by its timestamp, e.g., "Frame at 2.375s:". Please find one or more time intervals where the impact on a person's body described by the sentence "{query}" is physically applied, determining their starting and ending times. The query may match multiple separate impact moments in the same video. Return all matching intervals, not just the most salient one. Localize only the moment when the impact is actually applied to the person's body, not the approach or aftermath. Answer in the format "The event happens in x1 - y1 seconds; x2 - y2 seconds; ...", where each xi is a start time, and each yi is an end time, with xi $<$ yi. Return all matching moments in chronological order.

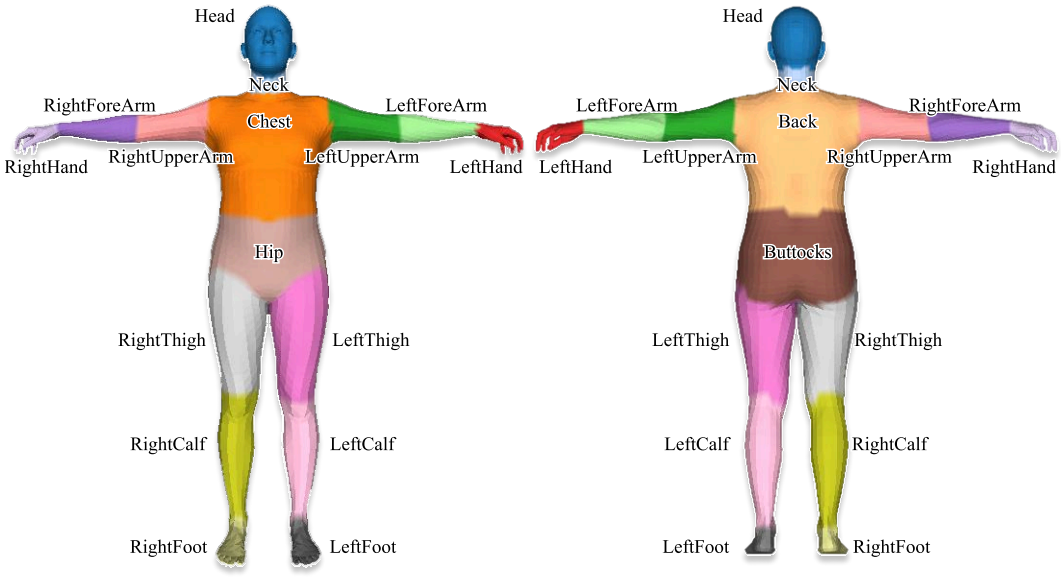


Figure 13: The 18 body-part labels for query-guided region selection. Left: front; right: back.

For generation-based baselines, the input video was provided either as uniformly down-sampled frames or as a raw video file, depending on the model interface. For frame-based inputs, the recommended frame sampling setting of each model was followed when available. Specifically, VTimeLLM, TRACE, and Grounded-Video-LLM were given 100, 128, and 96 sampled frames, respectively, while the remaining models received up to 192 frames. For videos shorter than the target number of sampled frames, input frames were limited by the video length. Each model was prompted to output one or more temporal intervals corresponding to the queried impact event. Because the baselines follow different input conventions, we adapt the prompt to each model's native interface while keeping the impact-localization instruction consistent. Table 7 lists the resulting prompts. We used deterministic settings for all open-source models, but evaluated each closed-source MLLM over five trials and reported the mean with standard deviation.

A.6.3 Spatial Localization Evaluation

We select BSTRO [27], DECO [64], and GraphiContact [41] as baselines. We exclude methods that require auxiliary context beyond human and scene observations, such as object meshes or hand-held object annotations, since such inputs are not consistently available in movie clips.

Spatial evaluation was conducted on each frame within the annotated temporal intervals of the queried impact event. For human-scene contact baselines, models were fine-tuned from the officially released pretrained weights. BSTRO was initialized from RICH-pretrained weights and fine-tuned with a learning rate of 5×10^{-5}. DECO and GraphiContact were initialized from DAMON-pretrained weights and fine-tuned with learning rates of 5×10^{-6} and 8×10^{-3}, respectively. DECO was trained in its no-context setting for image-only evaluation, while GraphiContact was fine-tuned without its semantic scene and body-part segmentation auxiliary losses. All three models were trained for 100 epochs with checkpoint selection on the validation split. For baselines whose released checkpoints predict only on the SMPL topology, we converted our SMPL-X annotations to SMPL before training and evaluation.

For Cosmos-Reason2, spatial localization was reformulated as query-guided region selection. The candidate regions were defined by extending the body-part segmentation of Chen et al. [11] with the neck region from the official SMPL-X body segmentation, resulting in 18 body-part labels. Figure 13 shows these part labels as color-coded regions on the SMPL-X mesh. Since this reformulation predicts body-part labels rather than vertices, we further report label-level F1 scores, with the results summarized in Figure 6.

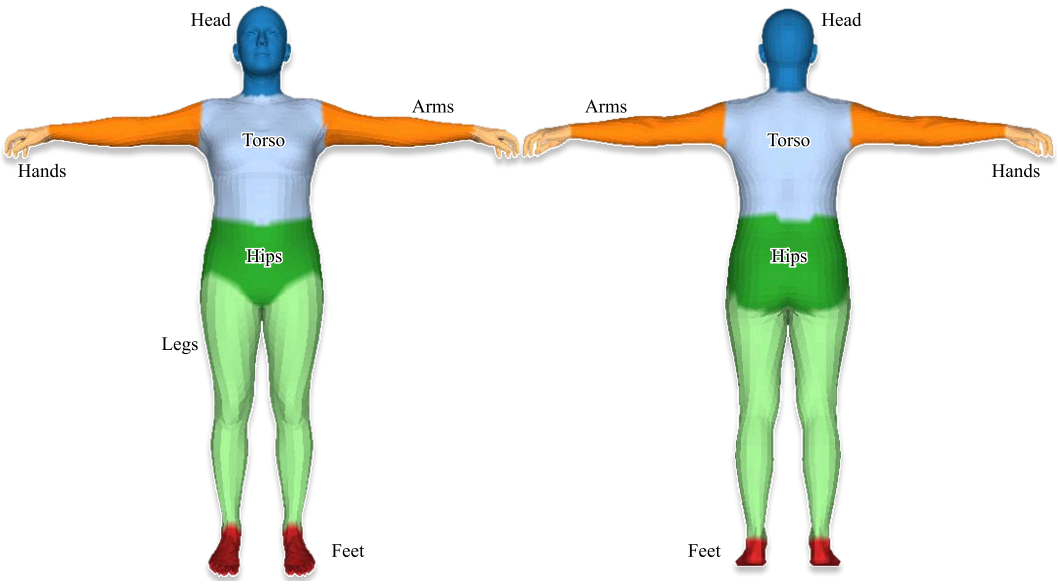


Figure 14: The 7 body-region groups for per-region evaluation. Left: front; right: back.

Both the 2B and 8B variants were fine-tuned using the HuggingFace implementation with LoRA. The LoRA setup followed rank $r = 32$, $\alpha = 32$, dropout 0.05, a learning rate of 5×10^{-5}, and 50 warmup steps. Both variants were trained for 3 epochs, and all evaluations used deterministic settings. The input-output format is provided below:

Prompt for Query-Guided Region Selection

You are observing a movie frame. The impact query is "{query}". Select all body regions physically affected by the impact in this frame.

The possible body regions are: head, neck, chest, back, left upper arm, left forearm, left hand, right upper arm, right forearm, right hand, butt, hip, left thigh, left calf, left foot, right thigh, right calf, right foot.

Answer with only the selected body-region labels, separated by commas.

Answer Format

The affected body regions are {list of body-region labels}.

To ensure consistency across all baselines, precision, recall, F1, and geodesic error were computed in the same SMPL space, with the results reported in Table 4. We further report F1 scores separately for each body region after merging the body-part labels into 7 regions. For each region, F1 was computed only on frames where that region was affected in the ground truth, using the vertices belonging to that region. The final score was averaged over those selected frames. Figure 14 shows the region groups.

A.6.4 Ablation Settings

Effect of temporal sampling interval. We first ablated the temporal sampling interval for retrieval-based baselines. Moment-DETR and QD-DETR were trained on the HAPACT training split and evaluated on the test split under different sampling intervals, as shown in Table 8. For intervals of 0.5s or longer, models used both CLIP and SlowFast features, while intervals shorter than 0.5s used only CLIP features. Overall, denser sampling improved temporal localization performance, with the 0.05s interval achieving the competitive results. Since human-centric impact localization requires fine-grained, frame-level temporal supervision, we adopted the 0.05s sampling interval for the main retrieval-based evaluations.

Table 8: **Effect of Temporal Sampling Interval.** Moment-DETR and QD-DETR are trained on the HAPACT training split and evaluated on the test split under different temporal sampling intervals.

Method	Sample Int.	mAP@0.3	R1@0.3	mAP				mR@k		
				G-mAP	@1_tgt	@2_tgt	@3+tgt	@1	@2	@3
Moment-DETR	2.0s	15.65	12.70	16.54	17.44	16.66	6.99	13.67	14.91	12.87
	1.0s	10.29	9.93	11.89	13.07	11.31	8.98	11.96	10.54	6.51
	0.5s	11.12	12.01	11.40	12.10	11.27	8.01	12.75	11.03	8.99
	0.1s*	16.98	13.86	18.65	19.82	17.30	12.22	14.60	16.00	19.38
	0.05s*	16.25	12.24	17.70	17.19	20.11	13.25	13.63	18.85	18.14
QD-DETR	2.0s	16.60	11.32	19.68	20.05	21.14	10.31	14.60	18.97	16.90
	1.0s	13.67	9.24	16.25	15.67	18.90	9.50	12.05	15.64	16.90
	0.5s	17.51	14.09	20.13	21.05	20.21	11.91	17.00	18.30	17.83
	0.1s*	16.02	12.93	18.84	18.94	20.63	8.74	15.43	19.40	15.66
	0.05s*	20.06	16.63	21.36	22.36	21.01	12.83	18.20	20.67	18.29

* Models trained with only CLIP features (intervals $< 0.5s$ are too short to extract SlowFast features).

Table 9: **Effect of Pre-training and Fine-tuning.** Retrieval-based models implemented in the Lighthouse [48] codebase are evaluated under three settings: QVHighlights pre-trained checkpoints without fine-tuning (*Zero-shot*), training from scratch on the training split (*HAPACT*), and fine-tuning the QVHighlights checkpoints on the training split (*Fine-tune*). All models use CLIP features with a temporal sampling interval of 0.05 seconds and are evaluated on the HAPACT test split.

Method	Setting	mAP@0.3	R1@0.3	mAP				mR@k		
				G-mAP	@1_tgt	@2_tgt	@3+tgt	@1	@2	@3
Moment-DETR	*Zero-shot*	8.05	5.77	10.04	11.69	7.07	5.56	6.93	5.88	7.29
	HAPACT	16.25	12.24	17.70	17.19	20.11	13.25	13.63	18.85	18.14
	Fine-tune	18.16	13.63	20.39	21.94	19.11	12.92	15.57	16.61	16.59
QD-DETR	*Zero-shot*	9.55	6.00	11.98	13.57	9.50	7.81	8.22	8.12	8.53
	HAPACT	20.06	16.63	21.36	22.36	21.01	12.83	18.20	20.67	18.29
	Fine-tune	21.27	16.17	22.06	23.91	19.98	15.39	16.86	18.67	19.69
EaTR	*Zero-shot*	9.49	6.24	12.28	12.85	11.78	6.29	8.13	9.64	11.01
	HAPACT	16.65	13.63	19.62	20.47	18.38	11.53	16.21	15.58	16.90
	Fine-tune	21.69	16.17	22.54	23.09	22.58	13.92	16.95	19.21	20.62
TR-DETR	*Zero-shot*	10.62	7.39	14.00	14.97	12.81	6.99	10.21	9.64	11.01
	HAPACT	13.62	10.16	14.87	15.78	13.91	9.36	11.46	11.76	12.40
	Fine-tune	20.86	16.17	22.20	23.86	20.36	14.62	17.69	17.94	21.39
CG-DETR	*Zero-shot*	7.96	4.62	11.09	13.05	8.27	5.08	7.11	6.06	7.29
	HAPACT	16.10	11.09	17.38	18.32	16.09	9.42	12.65	13.76	15.50
	Fine-tune	20.57	15.47	22.55	23.96	19.48	17.28	17.92	21.52	22.63
UVCOM	*Zero-shot*	10.20	7.39	13.65	15.56	10.71	5.98	10.67	8.12	8.22
	HAPACT	18.01	16.17	19.77	22.25	15.55	9.46	18.15	18.24	19.54
	Fine-tune	21.22	17.32	23.58	24.28	22.96	16.53	19.26	21.76	24.65

Table 10: **Effect of Input Modality in the MLLM Baseline.** Text-only (*T*), image-only (*I*), and text-image (*T+I*) inputs are compared under the same query-guided region selection setting. The table reports overall localization performance and per-region F1 over seven coarse body-region groups.

Method	Modality	F1 ↑	Precision ↑	Recall ↑	geo. (cm) ↓	F1 Scores per Body-part ↑						
						Head	Arms	Torso	Hips	Legs	Hands	Feet
Cosmos-Reason2-8B	*T*	15.98	14.03	39.36	37.84	20.90	13.90	19.68	24.35	7.66	12.29	8.13
	I	16.46	14.48	36.02	39.65	19.83	15.99	19.26	26.47	10.76	17.79	7.60
	T + I	18.59	16.17	44.11	33.31	21.21	15.92	20.01	29.95	12.59	17.87	8.71

Effect of pre-training and fine-tuning. We further examined whether retrieval-based temporal models benefit more from task-specific training alone or from adapting pretrained moment-retrieval checkpoints. As shown in Table 9, models initialized from QVHighlights and fine-tuned on HAPACT generally achieved stronger performance than models trained from scratch, indicating that prior moment-retrieval supervision provides a useful initialization for human-impact localization. Based on this result, the main retrieval-based temporal baselines were initialized from the pretrained checkpoints then fine-tuned on HAPACT.

Effect of modalities in the MLLM baseline. For spatial impact localization, we ablated the input modalities of the MLLM baseline by comparing text-only, image-only, and text-image settings, as shown in Table 10. The text-image setting achieved the strongest performance, suggesting that impact queries and visual evidence provide complementary cues. Accordingly, the main MLLM baseline incorporates both modalities.

A.7 More Benchmark Results and Analyses

A.7.1 Analysis of Temporal Bottlenecks in VMR Models

Results in Table 2 and Table 3 suggest different performance patterns across model families. To better understand the temporal bottlenecks of each family, we conduct subgroup analyses by impact length and ground-truth moment count per query. We evaluate a strong representative from each model family: Sim-DETR, UniTime, Qwen3-VL-8B, and GPT-5.4, and summarize their subgroup performance in Table 11. Sim-DETR and Qwen3-VL-8B are strongly affected by shorter impacts, whereas GPT-5.4 shows a smaller relative decline. Multiple-moment queries lead to substantially larger drops for Qwen3-VL-8B and GPT-5.4 than for Sim-DETR. UniTime remains low across both analyses. These results show different sensitivities to impact length and the number of moments across the evaluated VMR models, while other temporal bottlenecks may also remain.

A.7.2 HSC Baseline Performance under Different Training Settings

To further evaluate whether HSC architectures can learn impact localization from task-specific supervision alone, we additionally train DECO and GraphiContact from scratch on HAPACT. BSTRO is excluded because its released implementation requires initialization weights that are not publicly available. As shown in Table 12, training from scratch does not consistently improve F1 score over fine-tuning across the evaluated models. The limited transfer of HSC models therefore cannot be attributed solely to their original pretraining and may also reflect limitations in their architectures or learning formulations for impact localization.

A.7.3 Effect of Region-Label Imbalance on Spatial Performance

Figure 6 shows that our MLLM baseline performs unevenly across body regions, with higher F1 scores on the head, back, and buttocks. The impact occurrence distribution in Figure 11 is also concentrated in several of these regions across all splits, motivating us to examine whether region-label imbalance contributes to this performance pattern.

We therefore construct a region-balanced HAPACT training set. Specifically, the original training set is minimally pruned to balance occurrence counts across the 18 body-region labels. After pruning, the occurrence count of each label differs from that of the rarest label by at most two. We additionally construct a randomly down-sampled set with the same number of frames as a control.

The baseline is then trained separately on each training set and evaluated on the original test set, with the results reported in Table 13. The region-balanced set improves mean region F1 while reducing both the Region Gap and Across-Region SD. By contrast, random down-sampling does not reduce the gap. These results suggest that the original region-label imbalance may have partly contributed to the baseline's lower overall performance, but does not fully explain the disparity.

Table 11: **Effects of Impact Duration and Moment Count on Temporal Localization.** Values in parentheses denote relative changes from the reference group, with darker red indicating larger decreases. For GPT-5.4, results are reported as mean ± standard deviation over five runs, with relative changes computed from the mean. **Bold** indicates the smallest decrease for each comparison.

(a) Impact Duration *(b) Moment Count*

Method	mAP@0.3 (%)			mAP@0.3 (%)	
	Long (7+f, ref.)	Middle (3–6f)	Short (1–2f)	Single (1, ref.)	Multiple (2+)
Sim-DETR	40.22	24.71 (-38.6%)	10.17 (-74.7%)	26.58	24.23 **(-8.8%)**
UniTime	27.31	1.04 (-96.2%)	0.00 (-100.0%)	16.42	5.90 (-64.1%)
Qwen3-VL-8B	46.92	26.81 (-42.9%)	3.22 (-93.1%)	28.31	15.04 (-46.9%)
GPT-5.4	51.32 ± 1.38	44.60 ± 1.76 **(-13.1%)**	25.72 ± 1.85 **(-49.9%)**	53.90 ± 1.38	23.58 ± 0.39 (-56.3%)

Table 12: **HSC Baseline Performance under Different Training Settings.** *Zero-shot* denotes evaluation using the released DAMON checkpoints, *HAPACT* denotes training from scratch on HAPACT, and *Fine-tune* denotes fine-tuning the DAMON checkpoints on HAPACT.

Method	Setting	F1 (%) ↑	Precision (%) ↑	Recall (%) ↑	geo. (cm) ↓
DECO	*Zero-shot*	8.28	9.87	11.84	83.84
	HAPACT	5.05	68.92	6.05	55.75
	Fine-tune	5.41	68.16	6.18	50.34
GraphiContact	*Zero-shot*	7.88	8.85	10.23	87.42
	HAPACT	7.74	8.71	10.10	88.11
	Fine-tune	7.34	30.12	14.18	50.02

Table 13: **Effect of Region-Label Imbalance on the Spatial Performance of the MLLM baseline.** Mean Region F1 is the mean F1 across all 18 body-region labels. Head/Torso comprises the head, neck, chest, back, hip, and buttocks, and the remaining 12 labels form Limb. Region Gap is the difference between the mean F1 scores of the Head/Torso and Limb groups. Across-Region SD is the standard deviation of F1 scores across all regions. Lower disparity values indicate more uniform performance. **Bold** indicates the best result.

Training Set	Region-level F1 (%) ↑			Disparity ↓	
	Mean Region	Head/Torso	Limb	Region Gap	Across-Region SD
Original (HAPACT)	36.57	49.54	30.09	19.45	12.63
Random Down-Sampling	37.25	**50.37**	30.68	19.69	13.54
Region-Balanced Down-Sampling	**38.04**	49.05	**32.54**	**16.51**	**11.51**

A.8 Dataset Examples

This section presents representative examples from the HAPACT training and validation sets to illustrate the structure of each video-query instance and its annotations. Each example consists of a movie clip, a textual query, one or more temporal intervals, and frame-level on-body impact labels. An annotation visualization video for these examples is included in the supplementary material.



Figure 15: **Examples of HAPACT.** The examples show three representative impact categories: (a) human-to-human contact (e.g., grabbing), (b) human-to-environment contact (e.g., being thrown against a surface), and (c) impact from external forces (e.g., blast explosions).

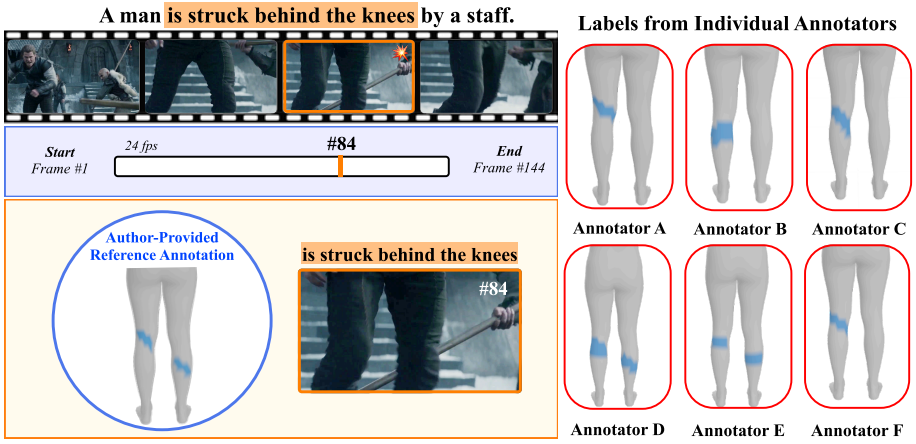


Figure 16: **Annotation subjectivity in spatial localization.** The reference annotation and individual annotator labels for the same impact frame and query show similar spatial tendencies, while precise vertex labels differ due to occlusion, narrow contact, and annotator interpretation.

A.9 Limitations

Annotation Subjectivity. Since HAPACT relies on human annotations, its labels may reflect how annotators interpret physical impacts from visual evidence and prior knowledge. This limitation is particularly relevant for spatial localization, especially when the impact is brief and affects a small body region. Figure 16 illustrates this ambiguity. For the query "A man is struck behind the knees by a staff" at frame 84, the author-provided annotation and six individual annotator annotations show similar overall tendencies, but differ in the precise annotated regions due to occlusion, the narrow contact area, and annotator-specific interpretations of the impact. Similar challenges have been noted by Tripathi et al. [64], where contact from images is inherently difficult and often requires human reasoning about occluded body-object relations.

Such ambiguity is a common limitation in human-annotated physical-interaction datasets, particularly at fine granularity. As reported in Section A.4.5, HAPACT achieves Fleiss' κ values of 0.53 at the vertex level and 0.61 at the body-part level. Related datasets report similar agreement levels: DAMON [64] reports $\kappa = 0.656$ for manually annotated SMPL contact vertices, and EPIC-Contact [4] reports $\kappa = 0.61$ over $3{,}106$ MANO vertices. Beyond agreement, HAPACT further applies quality-control procedures to maintain quality despite such variability, including annotator pre-screening and meta-annotator review, as detailed in Section A.4.5.

The resulting labels should not be interpreted as perfectly objective physical measurements. Instead, they provide a reliable source for learning the visual and semantic tendencies of human impact localization. Exact vertex selections may remain ambiguous in challenging cases, even when annotators agree on the affected body region. This limitation should be considered for precise vertex-level localization, whereas HAPACT remains well suited to applications where coarse spatial granularity is sufficient, including media-based haptic authoring and haptic experience generation.

Model-Lineage Bias. Gemini 2.5 Pro and GPT-4o were used during candidate generation and filtering, while the final labels were produced through human review and annotation. Although human review reduces direct model influence, lineage-specific biases introduced at these stages may remain. Models from the same lineages may therefore be better aligned with the data distribution, potentially reducing benchmark neutrality for GPT/Gemini-family models.

Limited Real-World Generalizability. HAPACT is constructed from movies, primarily motivated by media-based applications where audiovisual and narrative context is important. This focus may lead to domain shifts when applied to real-world videos, and HAPACT does not cover the full range of impacts, such as minor daily bumps, sports impacts, or medical-related collisions. At the same time, HAPACT covers diverse agents, objects, scenes, and impact events, as shown in Figure 3(f). The data construction pipeline itself is not specific to movies and could be extended to domain-specific videos, although such extensions and validation are left to future work.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction clearly state the paper's contributions: the human-centric impact localization task, the corresponding *HAPACT* dataset, and temporal and spatial benchmark evaluations. The claims are limited to physical impacts affecting a single target person, and the conclusions are based on the reported benchmark results and limitations discussed in the paper.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper discusses limitations in Section 4.3, including that the proposed spatial baseline is limited to coarse body-region prediction and does not yet provide fine-grained body-mesh localization. Section A.9 further discusses annotation subjectivity in spatial localization and describes quality-control steps to mitigate it.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

   Justification: The paper does not introduce theoretical results, formal theorems, or proofs. The main contributions are a dataset, task formulation, benchmark, and evaluation.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The paper describes the task definition, dataset construction, data splits, evaluation metrics, and baseline settings in Sections 3, 4.1, and 4.2. Section A.6 provides prompts and implementation details, while Section A.2 explains how the benchmark can be reproduced without redistributing the raw movie videos.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: Code and data are provided through anonymized GitHub and Hugging Face releases. The releases include annotations, metadata, splits, evaluation scripts, and instructions for reproducing the benchmark results.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: The paper specifies the training, validation, and test splits in Section 3.4, and describes the evaluated models, metrics, and benchmark settings in Section 4.2. Section A.6 provides implementation details, including feature sampling, prompts, training epochs, learning rates, checkpoint selection, and LoRA settings for the MLLM baseline.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

Justification: The paper reports mean and standard deviation over multiple runs for closed-source MLLMs to account for run-to-run variability.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Section A.6 reports the main hardware an configurations used for the experiments, including GPU resources and memory.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research conforms to the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: Section A.1 discusses potential positive applications and risks of applying impact prediction models beyond the intended context.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [Yes]

Justification: The paper describes safeguards for responsible dataset release. The underlying movie footage is not redistributed, and filtering procedure removes adult-rated movies and cases outside the task scope.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: Existing assets used in the paper are credited through citations and footnotes, including video sources, metadata sources, body models, annotation tools, and baseline models or codebases. Section A.2 states that the underlying movie footage remains owned by the original rights holders and is not redistributed.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: *HAPACT* is a new benchmark dataset with impact queries and temporal and spatial annotations. Its documentation is provided in Sections 3, A.5, and A.9. Corresponding documentation will be included with the released dataset.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [Yes]

Justification: Section A.4 provides annotator selection, compensation, task instructions, and the spatial annotation interface screenshot.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [Yes]

    Justification: Section A.4 describes IRB approval and the annotator risk disclosure.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: Sections A.3 and A.6 describe the LLM and multi-modal LLM usage in dataset construction, baseline design, and benchmark evaluation.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.