

---

# Supplementary Materials for HopaDIFF: Holistic-Partial Aware Fourier Conditioned Diffusion for Referring Human Action Segmentation in Multi-Person Scenarios

---

Anonymous Author(s)

Affiliation

Address

email

## 1 A Society Impact and Limitations

### 2 A.1 Society Impact

3 In this work, we introduce Referring Human Action Segmentation (RHAS), a novel task that enables  
4 fine-grained action understanding in complex, multi-person video scenes using natural language  
5 descriptions. We contribute the first dataset for this task, **RHAS133**, along with a diffusion-based  
6 referring human action segmentation solution, **HopaDIFF**, which integrates global and localized cues  
7 to achieve precise, target-aware action segmentation. These contributions lay important groundwork  
8 for the future of language-guided video understanding.

9 This research has the potential to significantly impact several real-world domains. In healthcare and  
10 eldercare, RHAS can enable systems to achieve action segmentation in shared spaces with multiple  
11 persons when a textual reference is given for a specific person, *e.g.*, “the elderly man in the red shirt”.  
12 In human-robot interaction, robots could better understand and track human behaviors in dynamic  
13 group settings. Similarly, in surveillance or safety monitoring, RHAS may help systems localize  
14 and interpret behaviors based on spoken or written descriptions, offering more intuitive and flexible  
15 interaction with video analytics tools.

16 However, this technology also raises ethical considerations. First, the reliance on data from publicly  
17 available movies may embed social, cultural, or gender biases present in media, which could affect  
18 model fairness and generalizability. For instance, actions associated with underrepresented groups  
19 may be poorly modeled, leading to biased or inaccurate outputs. Second, RHAS systems might be  
20 misused for invasive surveillance or profiling if deployed without safeguards, raising privacy and  
21 civil liberty concerns.

22 To mitigate these risks, future work should focus on curating more diverse, representative datasets and  
23 embedding fairness-aware training objectives. Transparency, stakeholder oversight, and responsible  
24 deployment frameworks will be critical to ensuring that the benefits of RHAS technologies are  
25 equitably distributed and ethically aligned with societal values.

#### 26 A.1.1 Limitations

27 As stated in our main paper, our work is the first to investigate human action segmentation guided by  
28 textual references to enable action segmentation for a specific individual in multi-person scenarios.  
29 Due to the novelty of the task, experiments are currently limited to our constructed dataset, which  
30 constrains the evaluation of the proposed model’s generalizability. Nevertheless, we made substantial  
31 efforts to ensure dataset diversity, collecting data from a wide range of movies over an annotation  
32 process that spanned more than 8 months. To further assess generalizability, we designed multiple

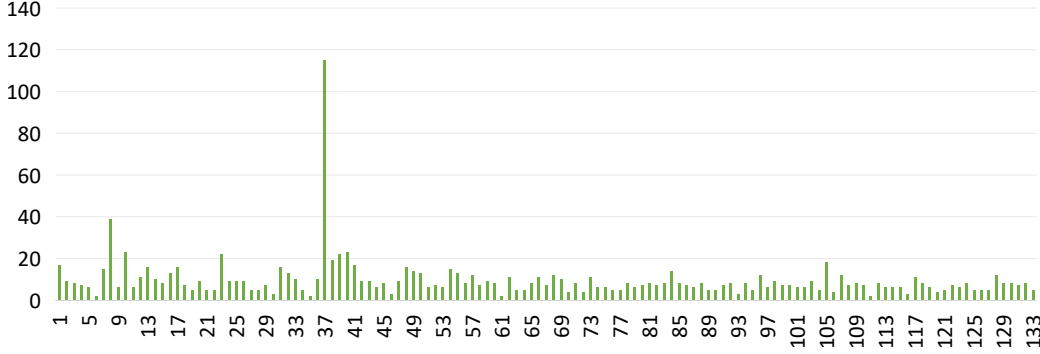


Figure 1: An overview of the statistics regarding the number of persons per movie in our **RHAS133** dataset. The horizontal axis denotes the video ID, and the vertical axis denotes the number of persons annotated in the corresponding video.

evaluation protocols, *e.g.*, random partition and cross-movie partition, and conducted cross-backbone experiments by replacing BLIP-2 [1] with CLIP. In future work, we aim to expand the dataset by incorporating a larger and more diverse set of movies.

## B More Implementation Details

In this section, we will provide more implementation details regarding our proposed HopaDIFF framework. For both encoders of the holistic branch and the partial branch, the number of feature maps is chosen as 64, and the kernel size is chosen as 5. We adopt a normal dropout rate of 0.5, a channel dropout rate of 0.5, and a temporal dropout rate of 0.5. For both the decoders of the holistic branch and the partial branch, we choose the dimension of time embedding as 512, the number of feature maps as 24, the kernel size as 5, and the dropout rate as 0.1. Regarding the hyperparameters of the diffusion process, we choose the timesteps as 1,000 and the sampling time steps as 25. The training epoch is set as 1,000.

## C Code, Dataset, and More Ablation Studies

Due to an urgent, unexpected maintenance of our GPU cluster during the supplementary preparation time, further ablation results are provided in our anonymous GitHub repository (<https://anonymous.4open.science/r/HopaDIFF-EF3D/>). The authors apologized for this inconvenience.

## D More Details of RHAS133 Dataset

In this section, we deliver more details of our contributed **RHAS133** dataset. First, we present the number of persons statistic for each movie in Fig. 1, which illustrates that our dataset contains diverse distribution regarding the number of persons per video.

## References

- [1] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*, 2023.