Chaotic World: A Large and Challenging Benchmark for Human Behavior Understanding in Chaotic Events (Supplementary Materials)

1. Details of Data Collection, Annotation, and Annotation Check

How we collect and process the videos. We first obtain the YouTube videos and process them into video clips with a resolution of 1280 (width) $\times$ 720 (height) at a frame rate of 25 frames per second.

How we annotate the video clips. Our dataset is carefully annotated by 15 annotators. Annotators are given instructions with reference videos on how to label the scenes, sounds, individuals, actions of individual and interactions between individuals. We use Label Studio [7] to annotate the videos. The video clips are viewed and annotated at five frames per second in Label Studio.

For our annotations, we identify scenes of interest, and initially label the event category based on the title or description of the video and then confirm it when carefully annotating the scene.

For the bounding boxes of the individuals, we follow AVA [2] style to label the bounding boxes.

For the Spatiotemporal Video Grounding, we further identify scenes of interest, their start and end time, and write descriptions of the scenes. The person is described based on, whichever that is applicable,

- the visible appearance of the person, *e.g.*
  - age and gender (*e.g.*, old man, young lady, girl),
  - hair color (*e.g.*, blonde)
  - style of clothes (*e.g.*, T-shirt, hoodie, shorts)
  - color of clothes (*e.g.*, standard colors such as black, red, blue)
  - accessories (*e.g.*, cap, spectacles, shades, watch, necklace, scarf, headdress)
  - color and/or type of bags (*e.g.*, black backpack), etc
- behavior(s) (*e.g.*, those outlined in Section 2.2),
- and interactions(s) with others (*e.g.*, those outlined in Section 2.3).

Given the large diversity of humans, behaviors, and interactions in our dataset, this gives rise to rich unique combinations of characteristics in different scenes. Furthermore, the vocabulary used for the description varies between annotators, and even for the same annotator. Thus, the 15 annotators generate very large diversity in the types of description.

For Spatiotemporal Action Localization and Behavior Graph Analysis, we create a dictionary to clearly define the various behaviors of individuals and interactions between individuals that can be useful to different groups of professionals (*e.g.*, emergency services, intelligence services, social behavioral scientists) in assessing and analyzing chaotic situations. This dictionary is constantly updated as we annotate the videos and further refined during the annotation quality check, where ambiguities are rectified and synonyms are standardized.

For Event Sound Source Localization, we use the same guidelines to annotate the bounding box for the visible object or person that is having/making the most prominent sound or voice in the scene. We also annotate the type of sound, as well as its start and end timing. The audio waveform of the video clip seen in the Label Studio's annotation tool allows us to assess when the sound/voice begins and ends.

How we check the annotations. For annotation quality checks, we conduct a total of two rounds, with different individuals cross-checking and verifying the correct start and end time of the video clips for Spatiotemporal Video Grounding task; bounding boxes of humans with the corresponding actions and interactions for Spatiotemporal Action Localization and Behavior Graph Analysis respectively; and bounding boxes of sound source for Sound Source Localization.

In Round 1, an independent annotator will check and update the annotations to ensure that there is no personally identifiable information or bias information (*e.g.*, sensitive racial profile) for the description of individuals, the bounding boxes are within the 8 pixel difference, the frames for the start and end timing are within 2 frames difference, and the descriptions of behaviors and interactions follow the definition outlined in the dictionary. In Round 2, a more experienced annotator will independently assess the annotations and ensure they meet the criteria stated above. Otherwise, 2 other annotators will disambiguate the annotation.

2. Different Levels of Details Crucial for Analyzing Chaotic Events

Our Chaotic World dataset provides annotations to different dimensions of a scene for holistic and detailed analysis of chaotic situations — from high-level details on the type of chaotic event (Section 2.1), to mid-level details of the complex interaction graph (Section 2.3) among individuals for Behavior Graph Analysis, and down to low-level details of actions (Section 2.2) of individuals for Spatiotemporal Action Localization, and sound and voice (Sec-

tion 2.4) for Event Sound Source Localization.

2.1. Various Chaotic Events

Our dataset contains a range of chaotic events such as:

- Natural Disasters/Phenomena (Earthquake)
- Natural Disasters/Phenomena (Flood)
- Natural Disasters/Phenomena (Hailstorm)
- Natural Disasters/Phenomena (Hurricane)
- Natural Disasters/Phenomena (Landslide)
- Natural Disasters/Phenomena (Tsunami)
- Natural Disasters/Phenomena (Volcanic eruption)
- Accidents
- Fires
- Queue chaos (*e.g.*, Shopping, Vaccination)
- Violence/Crime (Explosion)
- Violence/Crime (Fight)
- Violence/Crime (Quarrel)
- Violence/Crime (Gunshot/Terrorist attack)
- Protest/Riot

2.2. Wide Range of Actions

Our dataset contains a diverse range of actions and behaviors in chaotic events such as:

- aggressive behaviors, *e.g.*, fighting, hitting, kicking, pulling, punching, pushing, shooting/firing, throwing object
- destructive behaviors, *e.g.*, burning/setting fire
- criminal behaviors, *e.g.*, stealing/looting
- disruptive behaviors, *e.g.*, creating barricade, holding flarestick, spraying aerosol
- enforcement behaviors, *e.g.*, arresting, guarding, pinning
- life-saving behaviors, *e.g.*, performing resuscitation, carrying casualty, extinguishing fire
- movements, *e.g.*, kneeling, lying down, retreating, running/escaping, squatting, walking
- verbal behaviors, *e.g.*, arguing, cheering/chanting, praying, shouting
- non-aggressive behaviors, *e.g.*, clapping, dancing, holding flag, holding hands, holding signage, hugging, playing instrument, raising fist, watching, waving flag, waving flarestick
- media recordings, *e.g.*, recording, reporting live

2.3. Different Types of Interactions and Relationships

Our dataset contains a range of interactions and relationships in chaotic events such as:

- aggressive interactions (*e.g.*, fighting, kicking, pulling, pushing, throwing objects at),
- enforcement interactions (*e.g.*, arresting),
- verbal interactions (*e.g.*, arguing with, shouting at, speaking to),
- movement interactions (*e.g.*, following after, grabbing, holding onto, hugging, running away from, running towards, running with),
- non-aggressive behaviors (*e.g.*, watching),
- media recordings (*e.g.*, recording)

2.4. Diverse Types of Sound

Our dataset contains a range of sounds such as:

- voice (*e.g.*, arguing, booing, chanting/cheering, crying, shouting, shouting for help, singing, speaking, whistling),
- hand actions (*e.g.*, clapping),
- equipment (*e.g.*, airhorn, buzzing of walkie-talkie),
- materials and objects (*e.g.*, clanking (of metal), shattering, sizzling),
- firearms and explosion (*e.g.*, explosion, flare, gunshot/firing),
- musical instruments (*e.g.*, playing drums, playing musical instruments, trumpet),
- vehicles (*e.g.*, crashing, honking, motorcycle engine, siren (ambulance), siren (fire engine), siren (police), truck engine).

2.5. Diverse Range of Description of the Scene, Behaviors of Individuals and Interactions between Individuals

The large diversity in the types of description is generated from the combination of scene, description of the person's appearance, behaviors and interactions, as outlined in Section 1.

3. Additional Description of the Evaluation Metrics for the Tasks

The specific details and calculation of the evaluation metrics are as follows:

3.1. Spatiotemporal Action Localization

mAP refers to the mean Average Precision, whereby the Average Precision (AP) for each class is computed using $AP = \frac{1}{N}\sum_{s=1}^{s} Precision(s)$ and then averaged across all classes. Here N refers to the number of positive samples in the test set, while $Precision(s)$ refers to the precision for top s test samples. $Spatial\text{-}IoU$ refers to spatial Intersection-over-Union (IoU), which is defined as $\frac{Area_I}{Area_U}$, where $Area_I$ is the area whereby the predicted bounding box overlaps (*i.e.*, intersection) with the groundtruth bounding box, and $Area_U$ is the combined area (*i.e.*, union) of both the predicted and groundtruth bounding boxes.

3.2. Behavior Graph Analysis

$Recall@k$ ($R@k$) is the standard recall metric where $Recall = \frac{TruePositive}{TruePositive+FalseNegative}$ and k refers to top-k returned results. The prediction of interaction is considered to be correct only if the predicted tubelets (*i.e.*, bounding box of each individual tracked over time) of both individuals have an IoU of at least 0.5 with the groundtruth tubelets.

While there is another evaluation metric (*i.e.*, Graph Classification (GCls)) that is used in [6, 1] to evaluate the

model's performance in first identifying the classes of the bounding boxes (*e.g.*, person and various objects), and then predicting the relationship labels between bounding boxes. We, however, did not use this metric since our bounding boxes all belong to the same class of 'person', and evaluating GCls is the same as evaluating PredCls in our case.

3.3. Spatiotemporal Event Grounding

As described in the main paper, to evaluate the temporal grounding only, we use mean $temporal\text{-}IoU$ (m-$tIoU$). The m-$tIoU$ is computed by taking the average of $tIoU$ of all videos. $tIoU$ is defined as $\frac{F_I}{F_U} = \frac{F_{Pred} \cap F_{GT}}{F_{Pred} \cup F_{GT}}$, where F_I (or F_U) comprises the set of frames in the intersection (or union) of the predicted and groundtruth timestamps, and F_{Pred} and F_{GT} comprises the set of predicted ($Pred$) frames and groundtruth (GT) frames respectively.

To evaluate both spatial and temporal grounding (*i.e.*, Spatiotemporal Event Grounding), we use $spatiotemporal\text{-}IoU$ ($vIoU$) which is defined as $vIoU = \frac{1}{F_U} \sum_{t \in F_I} IoU(\hat{b}_t, b_t)$, where F_I (or F_U) comprises the set of frames in the intersection (or union) of the predicted and groundtruth timestamps, with $\hat{b}_t$ and b_t representing the predicted and groundtruth bounding boxes at time t respectively. $vIoU@\tau$ represents the percentage of samples whereby $vIoU \geq IoU$ threshold τ. The average of $vIoU$ of all videos (*i.e.*, mean spatiotemporal-IoU (m-$vIoU$)) is also computed.

3.4. Event Sound Source Localization

In line with [3], we employ the consensus Intersection over Union ($cIoU$). The groundtruth $Y_{i,j}$ is given by $Y_{i,j} = \sum_{k=1}^{A} \frac{Y_{k,i,j}}{A}$, where A refers to the number of annotators and $Y_{k,i,j}$ is defined as the groundtruth by k-th annotator, which takes the value of 1 if the bounding box includes (i, j)-th pixel, otherwise it takes the value of 0. A higher $cIoU$ score indicates better performance.

The consensus IoU is defined as

$$cIoU = \frac{Prediction \cap Groundtruth}{Prediction \cup Groundtruth} = \frac{\sum_{i,j} Y_{i,j} B_{i,j}}{\sum_{i,j} Y_{i,j} + \sum_{(i,j) \in \{(i,j)|Y_{i,j}=0\}} B_{i,j}} \quad (1)$$

where $B_{i,j}$ is obtained by binarizing the prediction $\hat{Y}_{i,j}$ using a threshold τ.

$$B_{i,j} = \begin{cases} 1 & \left(\hat{Y}_{i,j} > \tau\right) \\ 0 & \text{(otherwise)} \end{cases} \quad (2)$$

Area under the curve (AUC) is the total area under the probability curve that plots the True Positive Rate against False Positive Rate at various threshold values.

4. Additional Details of the Proposed IntelliCare Model

The following subsections describe additional details of the IntelliCare model with Dynamic Knowledge Pathfinder module.

4.1. Dynamic Knowledge Pathfinder

Dynamic Knowledge Pathfinder module automatically and intelligently chooses ('cares') the learning path in the network, with required task-shared and task-specific knowledge for each task. This module contains multiple dynamic layers, with each layer consisting of multiple knowledge blocks. We use I3D implemented with ResNet50 as the shared visual backbone and build the Dynamic Knowledge Pathfinder with it. We replace the last 5 layers of I3D (*i.e.*, layer4.1.conv2; layer4.1.conv3; layer4.2.conv1; layer4.2.conv2; and layer4.2.conv3) with dynamic layers.

4.2. Task Heads

Spatiotemporal Action Localization. The task head consists of a RoI feature extractor and one action classifier, analogous to the one used in [5].

Behavior Graph Analysis. The structure of the task head is equivalent to the original TRACE [6], which is composed of a hierarchical relation tree construction module, target-adaptive context aggregation module (to obtain the contextualized feature representation for each predicted relationship), and classification module.

Spatiotemporal Event Grounding. The task head has the similar structure as used in TubeDETR [8], which comprises a video-text encoder and a space-time decoder. In the video-text encoder, instead of using ResNet backbone, we use the shared I3D module to extract visual features from video clips.

Event Sound Source Localization. The task head is the same as the one used in [4], which consists of a predictive coding module (PCM) that iteratively predicts audio features based on visual features and then performs audio and visual feature alignment, and one attention module for calculating of the sound localization map.

4.3. Training and Testing

Training. During the forward propagation process, at each dynamic layer l, the visual feature f_v from the last layer is fed-forward to all the B blocks in the current layer to obtain the feature list $[f_v^1, f_v^2, ..., f_v^B]$. At the same time, the score module at this layer takes in the corresponding task indicator e_t, and outputs a one-hot vector S (with length $1 \times B$) to represent the score (0 or 1) of each knowledge block at this layer. Thereafter, the feature list is multiplied with the score vector S to yield the feature f_v^l of this layer. Thus, at each layer, the knowledge block with the

score of 1 is selected to be the best knowledge block containing the visual feature f_v^l for the current task.

In backward propagation, the weights of the selected knowledge block and the score module at each layer are both updated, and enables end-to-end training.

Testing. During testing, the parameters of the score module are fixed, and thus each task takes a fixed optimal path for inference. In other words, for each task, we select the most relevant block at each layer, and other irrelevant blocks do not need to be computed.

4.4. Visualization of Different Paths Selected for Different Task by the Dynamic Knowledge Pathfinder

The Dynamic Knowledge Pathfinder is trained to find the optimal block at each layer for each task. As shown in Fig. 1, different tasks will have different paths selected by the Dynamic Knowledge Pathfinder. In other words, each task will have a unique sequence of knowledge blocks to construct a unique path, which contains task-shared and task-specific knowledge. Some blocks will be shared among some tasks (*i.e.*, reused by multiple tasks), while some may only be used by one task. In Fig. 1, we can see that the Behavior Graph Analysis task and Spatiotemporal Event Grounding task share many blocks with the exception of layer 5. This could be due to that these two tasks share some similar knowledge of the visual scene for their respective tasks.

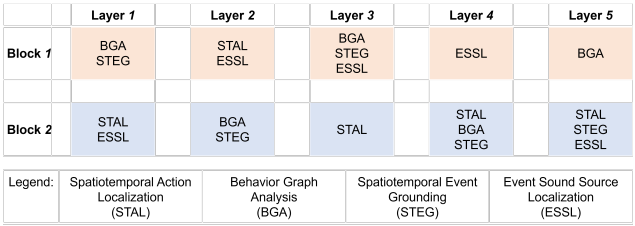

Figure 1. Example of tasks that use the block at each layer. This illustrates how different tasks may share the same knowledge block at each layer in the Dynamic Knowledge Pathfinder, and how each task will have a unique path — a unique sequence of best knowledge blocks for its task head.

References

[1] Yuren Cong, Wentong Liao, Hanno Ackermann, Bodo Rosenhahn, and Michael Ying Yang. Spatial-Temporal Transformer for Dynamic Scene Graph Generation. In *ICCV*, pages 16372–16382, 2021. 2

[2] Chunhui Gu, Chen Sun, David A Ross, Carl Vondrick, Caroline Pantofaru, Yeqing Li, Sudheendra Vijayanarasimhan, George Toderici, Susanna Ricco, Rahul Sukthankar, et al. AVA: A video dataset of spatio-temporally localized atomic visual actions. In *CVPR*, pages 6047–6056, 2018. 1

[3] Arda Senocak, Tae-Hyun Oh, Junsik Kim, Ming-Hsuan Yang, and In So Kweon. Learning to localize sound source in visual scenes. In *CVPR*, pages 4358–4366, 2018. 3

[4] Zengjie Song, Yuxi Wang, Junsong Fan, Tieniu Tan, and Zhaoxiang Zhang. Self-Supervised Predictive Learning: A Negative-Free Method for Sound Source Localization in Visual Scenes. In *CVPR*, pages 3222–3231, 2022. 3

[5] Chen Sun, Abhinav Shrivastava, Carl Vondrick, Kevin Murphy, Rahul Sukthankar, and Cordelia Schmid. Actor-Centric Relation Network. In *ECCV*, pages 318–334, 2018. 3

[6] Yao Teng, Limin Wang, Zhifeng Li, and Gangshan Wu. Target Adaptive Context Aggregation for Video Scene Graph Generation. In *ICCV*, pages 13688–13697, 2021. 2, 3

[7] Maxim Tkachenko, Mikhail Malyuk, Andrey Holmanyuk, and Nikolai Liubimov. Label Studio: Data labeling software, 2020-2022. Open source software available from https://github.com/heartexlabs/label-studio. 1

[8] Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. TubeDETR: Spatio-Temporal Video Grounding with Transformers. In *CVPR*, pages 16442–16453, 2022. 3