CSA-Graphs: A Privacy-Preserving Structural Dataset for Child Sexual Abuse Research

Supplementary Material

This supplementary material provides additional analyses and examples that complement the results presented in the main paper. In particular, we include qualitative examples of challenges in skeleton pose extraction observed in CSAI scenarios, as well as additional statistics and analyses of the scene graph representations in CSA-Graphs. These materials aim to provide further insights into the dataset's structural properties and to illustrate how the proposed representations capture contextual and pose-related information relevant to CSAI analysis.

A. Skeleton Pose Extraction

Challenges. Pose estimation may present limitations in certain CSAI scenarios. Some images contain only a single individual, which provides limited relational pose information for skeleton-based models. In other cases, the image depicts close-up views of specific body regions, such as genital areas, where reliable full-body pose estimation cannot be obtained. Figure 5 illustrates representative examples of these situations, highlighting cases where skeleton extraction is incomplete or unavailable.

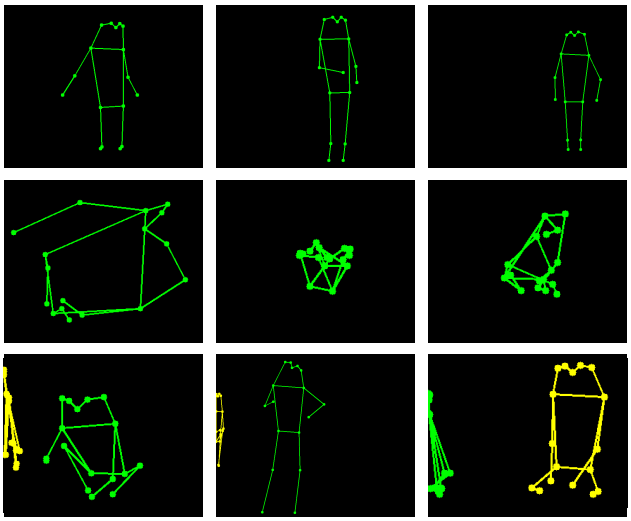
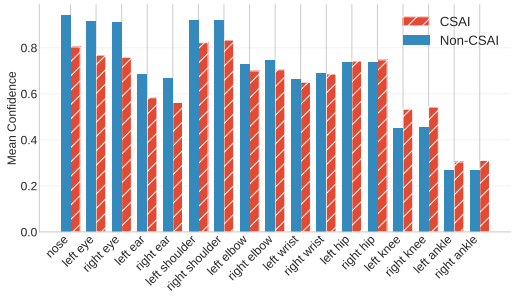


Figure 5. Examples of skeleton pose estimation limitations in CSAI scenarios. The top row shows images with only a single individual, resulting in a single detected skeleton and limited relational pose information. The middle row presents close-up views of specific body regions, where reliable full-body pose estimation cannot be obtained. The bottom row illustrates cases where skeleton extraction is incomplete due to occlusions, truncation, or challenging viewpoints.

Average Confidence Scores. As mentioned in Section 4.2.1, keypoints associated with the face as well as upper-body joints exhibit the highest detection rates across both CSAI and non-CSAI images. O the other side, distal joints such as wrists and ankles tend to be more frequently occluded or outside the image frame, leading to lower detection rates. A similar trend can be observed with higher average confidence scores (Figure 6). This correlation suggests that the pose estimator is more reliable for central body joints and facial landmarks.



(a) Mean confidence score per keypoint

Figure 6. Skeleton pose keypoint statistics in CSA-Graphs.

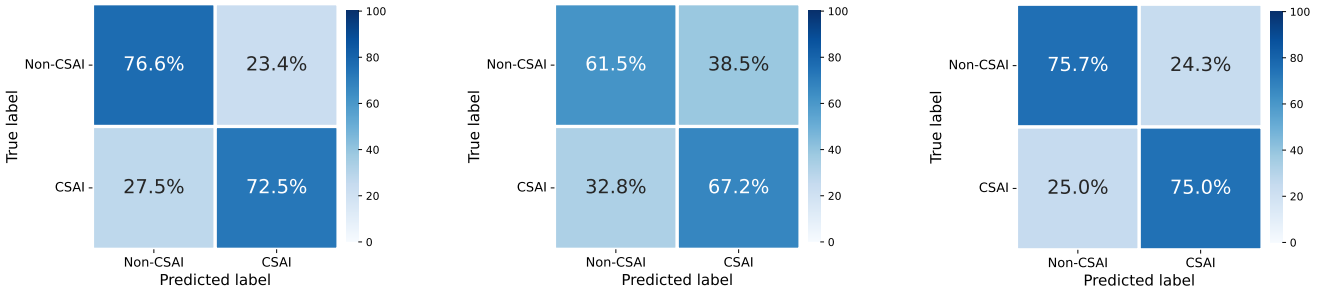
B. Confusion Matrices of Baseline Models

Figure 7 presents the confusion matrices of the baseline models evaluated in Section 5 of the main paper. The matrices provide a detailed view of classification outcomes for CSAI and Non-CSAI samples, showing the distribution of true positives, true negatives, false positives, and false negatives for each model. The results correspond to those obtained from experiments using 5-fold cross-validation.

As expected from the quantitative results reported in the main paper, the SG-baseline outperforms the skeleton-based model, reflecting the richer contextual information captured by scene graph representations. The Skl-baseline shows a higher number of misclassifications, consistent with the limitations of pose-only representations in certain CSAI scenarios. The fusion model reduces both false negatives and false positives compared to the individual modalities, further supporting the complementary nature of scene graph and skeleton pose representations.

C. Prediction Disagreement Analysis

Table 5 presents a prediction disagreement analysis between the skeleton-based model, the scene graph model,



(a) SG-baseline (scene graphs) (b) Skl-baseline (skeleton poses) (c) Fusion-baseline (both modalities)

Figure 7. Confusion matrices of the baseline models evaluated in CSA-Graphs.

and the fusion model. The results show that the two modalities capture complementary information about the scenes. The fusion model corrects a substantial portion of these errors: 143 cases in which the scene graph model fails and 372 cases in which the pose model fails are successfully recovered by the fusion approach. These results further support the complementary nature of the two structural representations and explain the improved performance observed when both modalities are combined.

Table 5. Prediction disagreement and correction analysis between skeleton-based, scene graph, and fusion models.

Prediction pattern	Samples	CSAI	Non-CSAI
Pose correct / Scene wrong	175	109	66
Scene correct / Pose wrong	350	137	213
Scene wrong / Fusion correct	143	86	57
Pose wrong / Fusion correct	372	162	210

Error Recovery Metrics. In addition to the disagreement analysis described above, we quantify how often the fusion model corrects errors produced by each modality. These measurements correspond to the recovery metrics reported in the main paper (Section 5.2) and provide a normalized view of the complementary behavior between the structural representations. Specifically, we define two metrics: the *pose error recovery rate* and the *scene error recovery rate*. These metrics measure the proportion of errors made by each single-modality model that the fusion model subsequently corrects.

Let E_{pose} denote the total number of samples misclassified by the skeleton-based model, and let $C_{\text{pose}\to\text{fusion}}$ denote the number of those errors that the fusion model correctly classifies. The pose error recovery rate is defined as:

$$R_{\text{pose}} = \frac{C_{\text{pose}\to\text{fusion}}}{E_{\text{pose}}}. \tag{1}$$

Similarly, let E_{scene} denote the total number of samples misclassified by the scene graph model, and let $C_{\text{scene}\to\text{fusion}}$ denote the number of those errors corrected by the fusion model. The scene error recovery rate is defined as:

$$R_{\text{scene}} = \frac{C_{\text{scene}\to\text{fusion}}}{E_{\text{scene}}}. \tag{2}$$

These metrics provide a complementary perspective to the disagreement analysis by measuring the relative proportion of errors that can be resolved through multimodal fusion. As discussed in Section 5.2 of the main paper, the fusion model recovers a substantial fraction of the errors made by the individual modalities, further supporting the complementary nature of the structural representations.

D. Frequent Scene Graph Triplets in CSAI and Non-CSAI Samples

Table 6 and Table 7 present the top-10 most frequent relational triplets observed in the scene graph representations for CSAI and Non-CSAI samples, respectively, ranked by the sum of confidence scores produced by the scene graph generator. Many of the most frequent triplets in both subsets correspond to anatomical relationships, such as *(hair, on, head)*, *(nose, on, face)*, and *(eye, on, face)*, which naturally arise from the frequent detection of human body parts in images containing people. However, differences emerge when considering contextual relationships involving people and surrounding objects. In the CSAI subset, triplets such as *(woman, laying on, bed)*, *(girl, laying on, bed)*, and *(man, laying on, bed)* appear among the most frequent relations, reflecting common spatial configurations present in the scenes. In contrast, the Non-CSAI subset contains several triplets associated with clothing, including *(boy, wearing, shirt)*, *(woman, wearing, shirt)*, and *(girl, wearing, shirt)*, which are absent from the top relations in CSAI samples. This difference is consistent with the observation that predicates related to clothing appear less frequently in CSAI images. Overall, these patterns illustrate how scene graph representations capture structural and contextual cues that may help differentiate CSAI from non-CSAI scenar-

ios while preserving a privacy-preserving abstraction of the original images.

Table 6. Top triplets in CSAI samples ranked by summed confidence score.

Triplet	Count	Sum score	Mean score
(hair, on, head)	1270	804.478	0.633
(nose, on, face)	862	594.013	0.689
(woman, has, hair)	685	403.372	0.589
(woman, laying on, bed)	369	369.000	1.000
(girl, has, hair)	627	352.883	0.563
(eye, on, face)	505	314.005	0.622
(girl, laying on, bed)	309	309.000	1.000
(ear, on, head)	400	251.421	0.629
(nose, on, head)	336	230.271	0.685
(man, laying on, bed)	218	218.000	1.000

Table 7. Top triplets in Non-CSAI samples ranked by summed confidence score.

Triplet	Count	Sum score	Mean score
(hair, on, head)	1569	1080.801	0.689
(nose, on, face)	1350	1019.411	0.755
(woman, has, hair)	998	563.683	0.565
(eye, on, face)	732	535.964	0.732
(ear, on, head)	693	500.922	0.723
(boy, wearing, shirt)	826	451.690	0.547
(girl, has, hair)	740	424.042	0.573
(woman, wearing, shirt)	933	392.642	0.421
(girl, wearing, shirt)	830	376.818	0.454
(nose, on, head)	436	305.143	0.700