Appendix

We first introduce the pose estimation pipeline in dataset construction in Appendix A. In Appendix B, we specify the details of the method implementation. We individually evaluate each module in our method, and the results are in Appendix C. We also show some qualitative results in Appendix D to demonstrate our method's capability of inferring the past. Appendix E is an ablation study on our method pipeline. Finally, we state the approval for our Thermal-IM dataset in Appendix F.

A. Pose estimation in dataset construction

We implement a two-stage pose extraction pipeline to acquire smooth and accurate 3D poses from RGB video pairs.

In the first stage, we use [4] to estimate a coarse 3D pose for every RGB frame. The resulting poses are in OpenPose [2] *body_25* skeleton with 25 joints. According to the pose estimation results, we divide the videos into continuous segments, ensuring that the character is always in the view of both cameras in each segment.

In the second stage, we synthesize the monocular pose estimation results of the two cameras to improve the pose quality. We first implement a triangulation step for more accurate depth estimation. Specifically, for each timestamp t, let p_t, q_t be the coarse 3D poses in the camera coordinates detected from the two cameras. We find a pair of scales a, b that minimizes the ℓ_2 distance between $p_t \cdot a$ and $q_t \cdot b$ in the world coordinate. After that, we use EasyMocap [1] to refine the pose sequence further. It smooths a sequence of 3D poses by optimizing the SMPL body parameters [6].

We further eliminate the failure cases of pose estimation. Specifically, all the poses are clustered into 2000 groups, and we manually filter out the clusters representing contorted poses.

B. Implementation details

Data processing: In our task, we use the first 15 joints out of the 25 joints in the OpenPose skeleton to represent a human pose. The first 15 joints are enough to depict human actions while ignoring the details such as ears and toes.

The input image size for our model is 288 × 384. And the evaluation metric MPJPE is also computed at this scale. The RGB and thermal lenses of our RGB-Thermal camera have different fields of view, and that of the thermal lens is smaller. We resize the thermal images in a preset way to align with the human poses, which are estimated in RGB image space.

Module	Learning rate	Batch size
GoalNet	5×10^{-5}	32
TypeNet	5×10^{-5}	128
PoseNet	1×10^{-4}	32
Semantic score model	3×10^{-5}	128

Table 1. **Learning rates and batch sizes.**

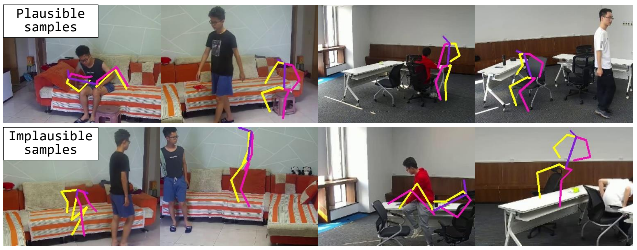

Figure 1. **Samples used to develop the semantic score classifier.** Plausible ones are samples in the dataset, while implausible ones are derived by random pose replacement, shift, and perturbation.

Model implementation: The backbones of GoalNet and PoseNet are both an Hourglass model [7] with three blocks, while that of TypeNet is ResNet18 [3]. The sizes of heatmap outputs of GoalNet and PoseNet are 72 × 96, and they are resized to be 288 × 384 by interpolation.

All modules are trained using the Adam optimizer [5] for 6k batch iterations. The learning rates and batch sizes are in Tab. 1. We use random crop and flip as data augmentation for all of them.

Semantic score: The data we use to train the semantic score model contains RGB images with plausible and implausible poses. Plausible poses are the 3s-ago poses, and implausible poses are derived by randomly replacing, shifting, and perturbing the plausible ones. Some samples are shown in Fig. 1.

Given an RGB image and a pose, we want a binary classifier to estimate how likely the pose is plausible. We use

Module	Average ℓ_2 Distance		
	Top 1	Top 3	Top 5
GoalNet	10.50	15.02	31.12

Table 2. **Evaluation of GoalNet.** We calculate the ℓ_2 distances from the top-1/3/5 predicted positions to the ground truth in the number of pixels.

Module	Accuracy		
	Top 1	Top 3	Top 5
TypeNet	10.50	15.02	31.12

Table 3. **Evaluation of TypeNet.** The task of TypeNet is indeed classification, so we evaluate the top-1/3/5 accuracy of its prediction.

ResNet18 as the model and train it with Binary Cross Entropy Loss. It is trained using the Adam optimizer with a weight decay of 1×10^{-3} for 6k batch iterations. The learning rate and batch size are in Tab. 1. Random crop and flip are used as data augmentation.

C. Individual evaluation of modules

As the three modules in our method are trained separately, we evaluate their performances in their own tasks in the following.

GoalNet: For each test instance, GoalNet samples 30 torso joint positions according to the predicted heatmap, and we evaluate how close they are to the ground truth 3s-ago position. We sort the 30 positions by order of their distances to the ground truth and compute the average ℓ_2 distance of the top-1/3/5 ones. We show the results in Tab. 2.

TypeNet: We evaluate TypeNet as a classifier and report its top-1/3/5 accuracy. The results are in Tab. 3.

PoseNet: We examine how the refinement of PoseNet makes an inputted pose type center closer to the ground truth 3s-ago pose. We report the MPJPE of poses before and after refinement in Tab. 4.

D. More qualitative results

In Fig. 3, we show samples of our method's synthesized poses in the test set. The involved indoor actions include sitting on a sofa/chair/table, lying on a sofa, touching a cabinet/bottle, and several actions on a yoga mat (sit-ups, push-ups, and leg stretching).

Module	MPJPE	
	Before	After
PoseNet	8.87	8.59

Table 4. **Evaluation of PoseNet.** Given the cluster center pose as input, we evaluate how much our PoseNet can refine it. The table shows the MPJPE from the poses to the ground truth poses before and after PoseNet refinement.

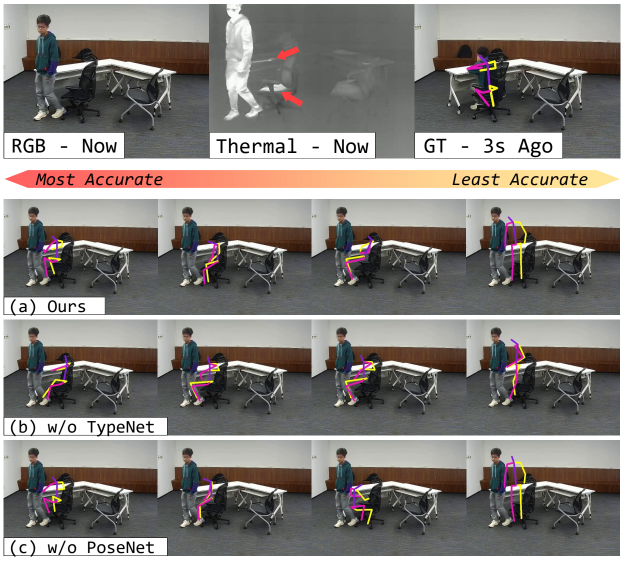

Figure 2. **Ablation study on model architecture.** From the thermal image, we can deduce that the person was sitting on the chair with arms on the table. In the model w/o TypeNet, predicted poses are often out-of-shape. The model w/o PoseNet can hardly provide the pose we desire because the number of pose types is limited. Our full model can refine the center pose of a type to fit with the details in the image, so it successfully generates sitting poses with an arm on the table (the 1st and 3rd column).

E. Ablation studies on pipeline modules

We implement two versions of our model without TypeNet or PoseNet to see how these modules contribute to our method.

w/o TypeNet: In a model without TypeNet, PoseNet generates a pose based on the input image and a root position given by GoalNet. The type of the synthesized pose is not specified here. In some cases, however, various poses are possible at a specific position. The skeleton joints generated by this model cannot be guaranteed to belong to the same pose, which leads to out-of-shape results as Fig. 2(b) shows and low semantic score in Tab. 5. Besides, because the generated poses are far from reality, the top-1 MPJPE is much higher than our complete model, though the top-5

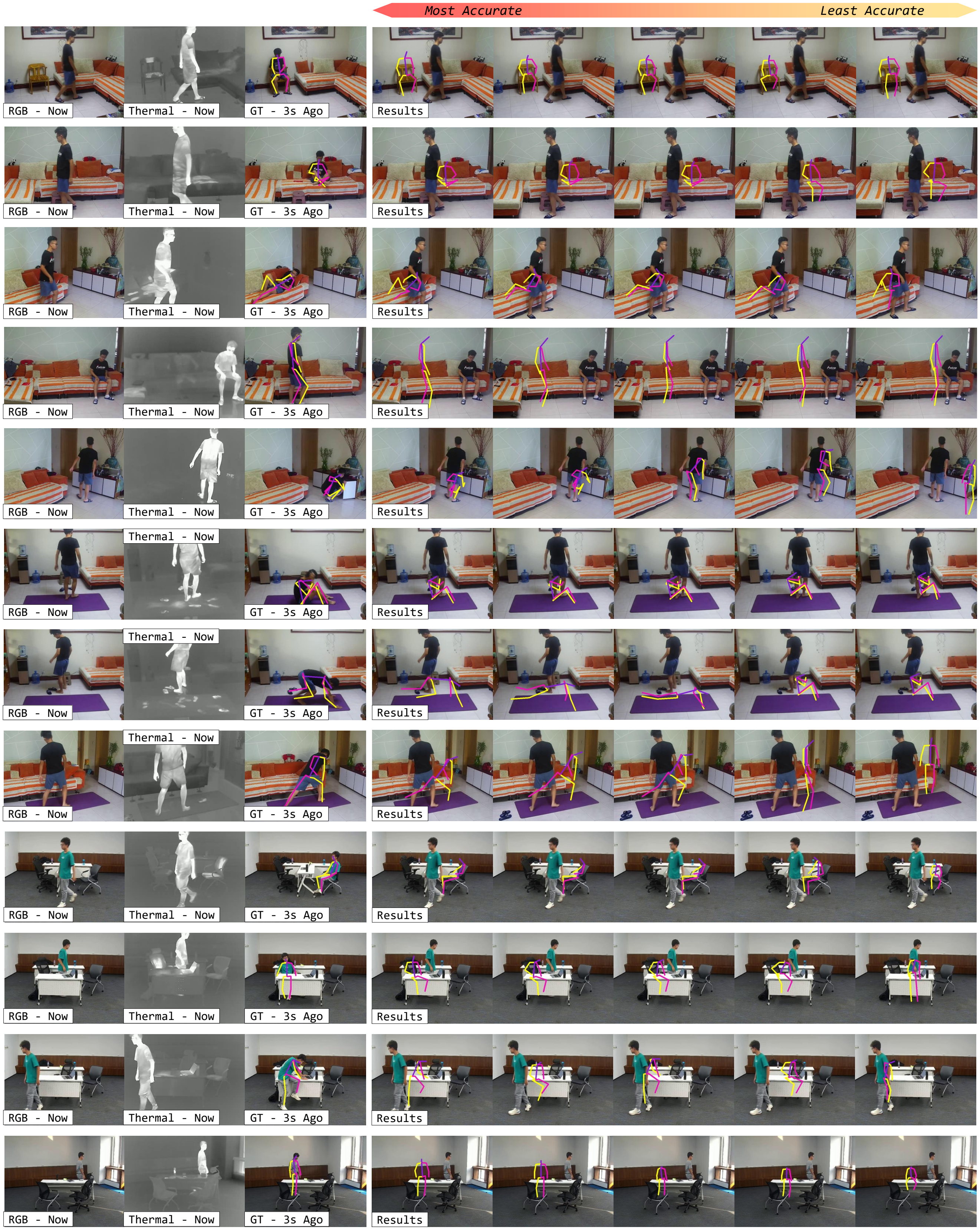

Figure 3. **Visualization results of our model.** For each sample, we sort the 30 predictions in order of MPJPE and show the 1st, 3rd, 5th, 10th, and 20th poses from left to right. Please pay attention to the bright marks pointed out by the arrows in the thermal images.

Modules			MPJPE			NLL	Semantic Score(%)
GoalNet	TypeNet	PoseNet	Top 1	Top 3	Top 5		
✓		✓	19.04	22.45	**25.19**	112.28	73.12
✓	✓		18.64	22.61	25.65	N/A	76.68
✓	✓	✓	**18.33**	**22.25**	25.25	**103.75**	**82.11**

Table 5. **Ablation study of removing different components.** Our model (the last row) outperforms the incomplete ones in most metrics, though removing TypeNet provides a slightly lower Top-5 MPJPE. We do not report the NLL for the one without PoseNet since it cannot be calculated in this setting.

MPJPE is competitive.

w/o PoseNet: In a model without PoseNet, TypeNet provides a pose type, and the center pose of this type is moved to the GoalNet's predicted position to serve as an answer. Since the number of pose types is limited, the duplicated pose cannot always fit with the details in the image. In the first column of Fig. 2(a) *vs.* (c), the model without PoseNet simply draws a sitting pose, while our complete model refines it so that the right arm is put on the table. As Tab. 5 illustrates, the refinement served by PoseNet improves both the synthesized poses' similarity to the answer and the plausibility in the context.

F. Approval

We have obtained approval for collecting and using the Thermal-IM dataset from the Institutional Review Board of our university department.

References

[1] Easymocap - make human motion capture easier. Github, 2021. 1

[2] Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019. 1

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, jun 2016. 1

[4] Hanbyul Joo, Natalia Neverova, and Andrea Vedaldi. Exemplar fine-tuning for 3d human model fitting towards in-the-wild 3d human pose estimation. In *International Conference on 3D Vision (3DV)*. IEEE, dec 2021. 1

[5] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. Dec. 2014. 1

[6] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM Trans. Graphics (Proc. SIGGRAPH Asia)*, 34(6):248:1–248:16, Oct. 2015. 1

[7] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *European Conference on Computer Vision (ECCV)*, pages 483–499. Springer International Publishing, 2016. 1