SCAIL: Towards Studio-Grade Character Animation via In-Context Learning of 3D-Consistent Pose Representations

Supplementary Material

A1. Details on Pose Conditioning

Rendering Details. To preserve spatial relationships between joints, we represent the human skeleton using cylindrical structures. For efficient rendering of multiple skeletons, we employ a 3D rendering pipeline that first converts the cylinders into spatial voxels, followed by ray marching implemented via [17]. This strategy is highly optimized for modern GPUs, introducing negligible computational overhead.

Augmentation Details. As our augmentation strategy is designed to maximally preserve the original motion, we use a high augmentation rate of 0.8 to achieve the balance between the pose-following accuracy and motion transfer robustness. During training, the overlaid 2D hand and face keypoints extracted by DWPose [47] will go through additional augmenation after the 3D adaptation process where 2D keypoints are shifted to match the reformed 3D skeleton after body rescaling and camera manipulation. This step is to minimize the unintended influence of the 2D facial and hand signals on the 3D pose representation.

Performance Tradeoff. For the injection of the representation in *full-context* settings, we observe that the rendered pose sequence is relatively sparse as most frame areas consist of non-informative black pixels. In our method, *spatial pooling* to pose sequence can serve as a simple workaround to effectively reduce the contextual pose tokens to $1/4$ and preserves accurate pose-following capability. In addition, *full-context pose injection* introduces no new parameters except the additional patchify layer to the original model, offering a more streamlined architecture compared to stacking additional DiT layers in residual context-tuning methods [18]. For quantitative comparison, we compare the inference costs of the two injection schemes (512 * 896 resolution, 81 frames, 20 diffusion steps) on an H100 GPU. In general, considering the significant motion error reduction, we conclude that the modest efficiency trade-off is acceptable, particularly in studio-grade scenarios which prioritize stringent accuracy and stability.

Methods (all w/ CFG)	Inference Time (s)	FPS	Memory (GB)
14B *w/* channel-concat	286.11	0.283	61.7
14B *w/* full-context (Ours)	380.78 (**+33.1%**)	0.213 (**-24.9%**)	68.5 (**+11.0%**)

A2. More Ablation Studies

We conduct further ablation studies on **SCAIL**-1.3B model to evaluate the contribution of our proposed components. All training settings, including learning rate, data, batch size, and training steps, are kept exactly the same across all models in our ablation studies. Additional user studies are conducted to collect users' preference towards different configurations for the **SCAIL**-1.3B model.

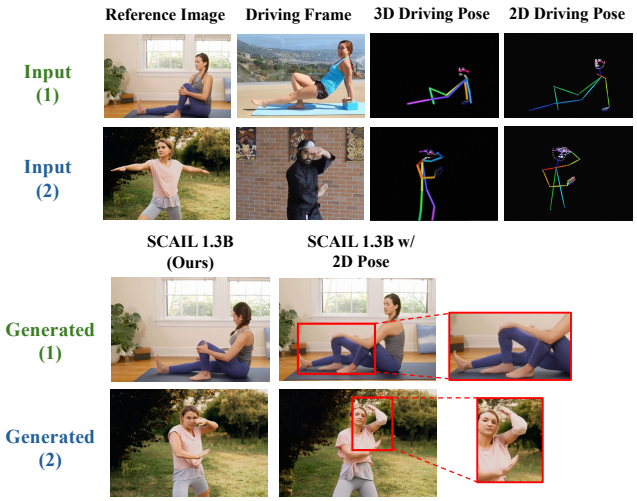

Figure A1. Ablation studies on pose representation. Anomalies in the human body or deviations from correct posture are boxed.

A2.1. Ablation details on the Pose Representation

As mentioned in the main paper, we implement the 2D augmentation strategy as close as to the 3D version, with similar figure settings and same augmentation ratio of 0.8. Faces and hands are also augmented identically to the 3D version and retargeting logic from [39] are applied during inference. As shown in Table 2, the estimation noise and the *2D-3D mismatch* noise introduced by the pipeline significantly undermine the model's performance in the self-driven subset of our challenging Studio-Bench, which involves a high ratio of complex motions.

For the cross-driven subset, 2D pose can easily lead to distorted limbs in generation especially when the 1.3B model have difficulty transferring motion to a significant different reference image, as seen in case (1) of Figure A1. Case (2) indicates that when model needs to distinguish the front and back of limbs, the inherent ambiguity of 2D pose can result in incorrect pose interpretations. Qualitative results from the cross-driven subset and quantitative results from the self-driven subset together demonstrate the overall effectiveness of our proposed 3D-based solution. Figure A2 can also demonstrate the unreasonable scaling factors introduced by 2D-based retargeting process during the inference

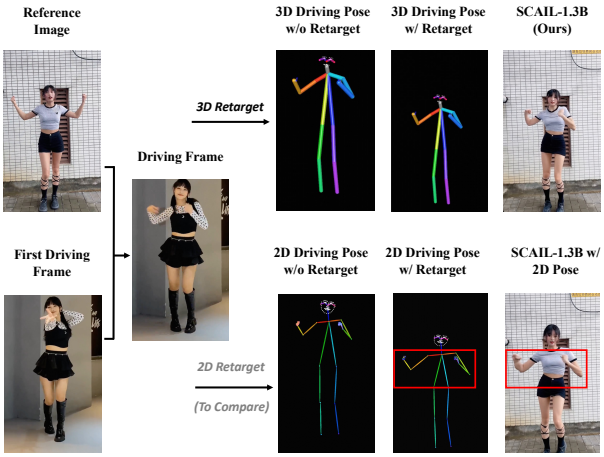

Figure A2. Ablation studies on pose retargeting. 2D Retarget are visualized for comparison. Regions where the body proportions deviate from the reference image are boxed.

step, which we will discuss in the next section of retarget ablations.

A2.2. Ablation on 3D-Consistent Adaptation

Furthermore, we validate the key component in our pose representation: 3D-Consistent Adaptation. 3D-Consistent Adaptation includes 3D Retarget in cross-driven inference and 3D Augmentation in training.

Ablation on 3D Retarget. Figure A2 shows the comparison of the driving pose with and without 3D Retarget. 3D Retarget helps transfer the motion to the person without introducing position change. Compared to 2D Retarget, 3D Retarget keeps the original motion without introducing limb length distortion. Note that in our **Studio-Bench**, we only include cases where 2D Retarget works well for a fair comparison of the the model itself's performance against other baselines. In wild scenarios, however, 3D information and camera parameters can help create highly robust retarget rules that are suitable for production-level use.

Ablation on 3D Augmentation. 3D augmentation is central to our method's ability to adapt to different characters. We conducted experiments to evaluate the impact of using 3D augmentation in the context of self-driven animation. The results indicated that even with a high augmentation ratio, there was no significant difference in the metrics compared to when augmentation was not used. This is because 3D augmentation effectively preserves motion information by only altering figure shape and maintaining the temporal motion semantics.

In the case of cross-driven animation, 3D augmentation significantly mitigated identity leakage, particularly notable in characters with substantial body shape differences, as illustrated in Figure A4. To quantify this improvement, we conduct a user study comparing two groups: with and without 3D augmentation. The results reveal that 3D augmentation clearly enhances the metric of *Physical Consistency* and *Identity Similarity*, endowing our model with better generalization capabilities when handling in-the-wild characters.

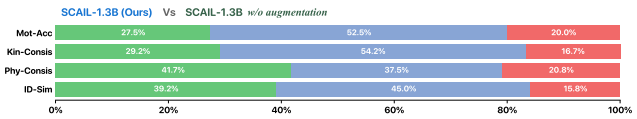

Figure A3. User study results of ablation on 3D Augmentation.

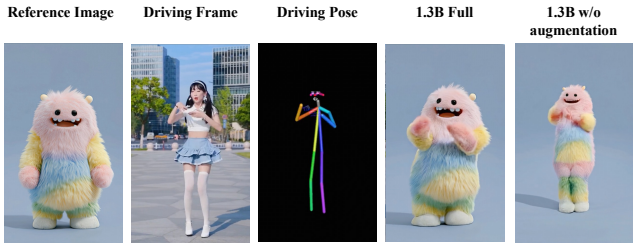

Figure A4. Qualitative results of ablation on 3D Augmentation.

A3. Data Source

Our training dataset is composed of three primary data sources: (1) samples retrieved from our internal base model training data and other downstream tasks, (2) a large collection of high-resolution dance videos from Bilibili and YouTube, and (3) additional sports videos such as gymnastics and figure skating. To ensure diversity, we maintain a certain proportion of stylized content, including 3D and 2D animations from source (1), as well as MMD and Live2D animations from source (2).

A4. Details on Studio-Bench

Benchmark Composition. To comprehensively evaluate studio-grade scenarios, we curate an diverse set of motion sequences in our **Studio-Bench**, as illustrated in Figure A5. The motion collection in our dataset emphasizes complex human-body configurations, covering a wide spectrum of challenging inputs, including dance, sports, martial-arts, acrobats and so on. In addition to isolated single-person motions, the test set also contains a small portion of interactions between the person and the environment, as well as several cases of multi-person interactions like dual dancing. We also include certain portion of fine-grained motions which are commonly featured in advertisement poses and iconic movie gestures, to evaluate the model's all-around capability.

For the construction of cross-driven cases, we intentionally select reference character images to cover diverse figure shapes and different facial characteristics. On top of these

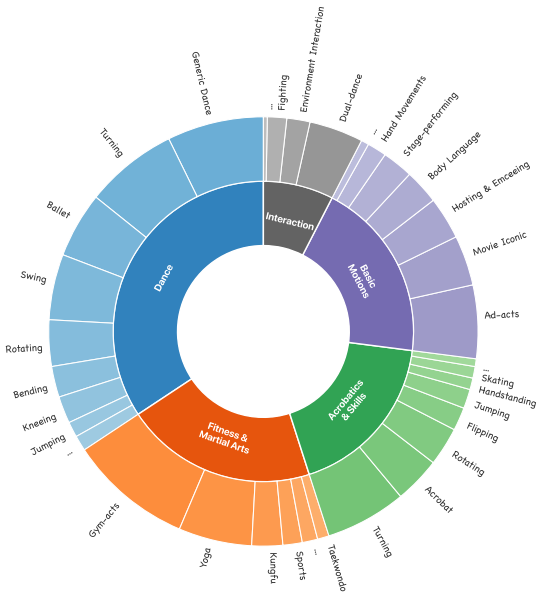

Figure A5. Visualization of the distribution of data annotations in **Studio-Bench**. We categorize and annotate videos based on motion types. The annotation of a single video can contain multiple tags, such as "turning" and "ballet".

real human references, we additionally introduce approximately 40 non-real characters. Most of them originate from 3D animated productions, while others include 2D animated characters, plush toys, anime figurines, and various stylized representations.

Metrics and evaluation details. For *self-driven* part we employed several widely-used quantitative metrics, including PSNR [15], SSIM [41], LPIPS [49], and FVD [33]. To evaluate the generated results in the second subset, we design four metrics: (1) *Motion Accuracy*, which measures how faithfully the generated motion follows the driving signal in a frame-by-frame manner. (2) *Kinesiology Consistency*, evaluating whether joint rotations and body movements remain anatomically feasible and temporally coherent, penalizing sudden twists or physically impossible poses that break natural motion continuity. (3) *Physical Consistency*, assessing whether the generated motions comply with basic physical constraints such as gravity, support, and momentum conservation, penalizing unrealistic behaviors like hovering in midair. (4) *Identity Similarity*, measuring the consistency of the subject's appearance with the reference image. To ensure methodological rigor, we note that the user study for selecting the best-performing model in Table 1 and the win/tie/lose evaluation in Figure 6 were conducted on different batches of participants, allowing the two sets of results to serve as cross-validations. All quantitative results on the *cross-driven* subset are based on the 120 single-character pairs to ensure fairness considering some baselines [36] are incompatible with multi-character settings. Evaluation samples are strictly excluded from the training data.

A4.1. More Examples on Studio-Bench

Example from the main paper demonstrate that our model can handle both fine-grained motions and highly complex in-the-wild motions. Figure A6 and Figure A7 will provide more cases with nonstandard figures to show our model's generalization ability across a wide range of subjects and artistic styles. Figure A6 demonstrates that SCAIL can produce accurate limb motions that respect the features of non-standard figures such as thin-limb anime characters. Furthermore, when the driving image is drastically different from the standard human figure (such as a plush toy with a very short body), SCAIL avoids the undesirable changes in body proportions that often plague baseline models. When the dual challenges of complex motions and non-standard character figures appear together in Figure A7, the issues with baseline models become even more pronounced. Another point worth noting is the scenario of *reverse driving*, where an anime character's motion is used to drive a real person. While most user inputs involve a real person driving another figure, this less common use case presents demand of making a real person mimic an anime posture. We found that previous methods produce strange proportions under such inputs in Figure A8 while our model is still capable of handling this task. These comparisons highlight the strong potential of our approach for studio-grade applications where character compatibility for diverse motion types are critical requirements.

A5. Discussion

A5.1. Limitations and Future Work

SCAIL currently relies on facial landmarks to achieve face control. Such a representation is inherently limited in fine-grained facial expression. As our work primarily targets addressing the challenges including instability and motion artifacts in studio-grade video generation, enhancing the expressiveness of facial control is left for future exploration. Specifically, future work will focus on improving the accuracy and fidelity of fine-grained details, such as hands and facial expressions, to further elevate the model's overall quality.

A5.2. Ethical Considerations

As our method advances character animation to a new level of realism and expressiveness, the potential for misuse, particularly in generating misleading or harmful digital content, becomes an important consideration. Despite these concerns, we believe that fully open-sourcing our model will bring substantial value to the community and encourage a wide range of responsible and creative works.

Figure A6. Comparison of model's ability to preserve body structure for non-standard character figures.

Figure A7. Visualization of our model's performance under both high-dynamic motion and non-standard character figures.

Figure A8. Visualization of our model's performance under *reverse driving* settings.