Holoported Characters – Rebuttal

We thank the reviewers for their valuable and overall positive feedback. In the following, we will refer to reviewer CjKF, Zmjn, J2A2 as **R1**, **R2**, **R3**, respectively.

Dataset and Details (**R1**, **R2**, **R3**). We will release code/data and add a section on data processing details.

Zoomed in Face (**R1**, **R3**). While the primary focus of this work is on full-body rendering, our results, e.g. video (7:25 to 7:40), demonstrate the ability of our method to capture both, hands and faces. As suggested, we will provide more zoomed-in results and comparisons as shown in Fig. a)

DDC (**R1**). DDC and our work have substantially different goals. DDC is a pose-driven avatar, i.e. learning a function from pose to geometry and appearance. In stark contrast, our work focuses on free-viewpoint rendering given sparse RGB views. We want to highlight that the DDC character model is merely a component in our pipeline. Our design answers the question: How to efficiently combine a learned human model (DDC) with the information present in the sparse view images to achieve real-time and high-quality renderings. Components such as the SRNet, Projective Texturing, and TexFeatNet are novel in this context, and our overall pipeline is unique in literature. We extensively ablate all design choices, significantly improve the previous state-of-the-art, and show unprecedented quality and performance. We believe our concepts and design will strongly benefit the community. The joint regression of appearance *and* geometry from sparse views (as suggested by **R1**) is an interesting direction for future work and we believe our work can provide a profound basis for this direction.

Evaluations (**R1**). We choose two subjects with different clothes from a well-established benchmark, i.e. trouser and dress, to test the robustness of our method. Here, the test set spans more than 7000 frames per subject showing highly varying poses. Thus, we believe the evaluation convincingly shows the effectiveness of our method. Moreover, we clarify that we indeed show qualitative comparisons *on more than one subject* (see Fig. 10 (supp mat) and video (6:33 to 6:38)). We agree that adding another subject would benefit our ablation studies. Therefore, we extend our ablation to subject *S2* under *novel pose synthesis* (see rebuttal table). Due to the limited rebuttal time, we can only provide qualitative comparisons on one new subject (see rebuttal figure b)). Results confirm the effectiveness of our approach. We will add a complete evaluation to the final version.

Limitations (**R1**). We discussed limitations in Sec. 1.9 (supp mat), and the core points will be included in the paper.

4K (**R1**). Tab. 2, Fig. 8, and L477-487 discuss the effects of using 4K resolution in our method. Further, we would like to point towards Fig. 11 in the supplemental (see insets of 4K and ours below) where the effect can be seen better. In the future, we plan to explore multi-resolution image encodings to preserve even higher frequency details.

Method/Metric	w/o Texture	w/o Features	w/o SR	**Ours**
PSNR↑	25.82	28.37	**28.42**	28.03
LPIPS↓ (×1000)	41.36	31.30	31.11	**28.49**
FID↓	55.17	21.05	20.85	**13.26**

Problem Setting (**R2**). The setting of one-time capture and sparse signal animation is important for immersive telepresence. The goal is that once someone is digitized, they should be able to drive a high-quality avatar from commodity hardware. This will enable applications in 3D video conferencing, immersive virtual talks, to only name a few. Our method is a step towards this goal, and we significantly improve over competitive baselines (DVA, which is the previous SOTA operating under the same setting) by demonstrating unprecedented photorealism and runtime. We agree that that changing the clothing is an aspect that our method currently does not address. However, we believe our work makes an important step towards this long-term goal.

Comparison with HDHumans (**R2**). We also highlight in the writing that those methods are solely pose-driven and, therefore, have separate discussions for image-based and pose-driven works. Nonetheless, the insight from such a comparison is that sparse image information can help disambiguate the task and, thus, lead to improved performance. We have further (and more extensively) compared to methods in the same setting as ours: 1) ENERF (that uses images as input), 2) DVA (that uses pose and images as input).

Incremental (**R2**). We politely disagree. Our work demonstrates how a 3D avatar prior and sparse image information can be efficiently combined for high quality free view rendering of humans in real time. While we leverage concepts from Graphics, we demonstrate a new way of combining explicit models with neural components. To the best of our knowledge, no one has presented such a design before. Moreover, our results clearly outperform prior works as also admitted by the reviewers. Thus, we believe there is merit from a technical and result point of view.

Misc (**R2**). 1) Flickering artifacts might be due to the DynaCap camera setting, which have slight differences in the color calibration. 2) Very fast motions might cause a degradation in tracking and, thus, cause visual artifacts. We will add a section in the limitation discussing this. 3) Flip comparison will be added to the video. 4) Normals are invariant to global translation. Our encoding assumes appearance is not changing with translations, it helps to prevent overfitting. 5) Rigging Details will be provided.

Temporal Stability (**R3**) of our method is extensively demonstrated/compared in the two supplemental videos.