

Feed-forward Gaussian Registration for Head Avatar Creation and Editing

Supplementary Material



Figure 1. MATCH predictions overlaid with texturized Gaussian splat renderings.

Contents. This supplementary material provides additional results, implementation details, and analyses that support the claims presented in the main paper. Section 1 gives more information on the datasets used for training and evaluation, followed by implementation details in Section 2. Section 3 provides more novel view synthesis experiments. Section 4 and Section 5 present reconstructions for in-the-wild and single-input-image scenarios respectively. Section 6 shows additional results for the interpolation, semantic editing, and expression transfer on Gaussian splat textures. Section 7 supplements a quantitative evaluation of the correspondences predicted by MATCH. Section 8 adds further comparisons of the subject-specific head avatars created from MATCH’s predictions. We close with ethical considerations in Section 9.

1. Datasets

Ava-256. The Ava-256 dataset [15] provides multi-view video captures from 80 cameras of 256 subjects performing a wide range of expressions. Ava-256 comes with ground truth mesh registrations of the head, including hair proxy geometry. We sample the captured videos at a framerate of 0.75 fps and select cameras that are evenly distributed over the frontal hemisphere in a range of $\pm 40^\circ$ horizontally and $[-15^\circ, +36^\circ]$ vertically. To facilitate generalization to other datasets with different backgrounds, we remove the image background using the provided alpha masks. We adopt Avat3r’s [9] train-validation split and use 244 iden-

titles for training and 11 for validation. One subject was removed from Avat3r’s validation set due to faulty ground truth segmentation masks. We sample 123,000 frames for training and 1,000 for validation.

NeRSemble. The NeRSemble v2 dataset [8] provides multi-view video captures from 16 cameras of 425 subjects, which are split into 419 training and 6 validation subjects. We select cameras

221501007, 222200045,
222200049, 222200043,
222200047, 222200038,
220700191, 222200041,
222200046, 222200040,
222200042, 222200044

for training and leave the rest for validation. As with Ava-256, we sample 123,000 frames for training and 1,000 for validation. Further, we use a pretrained matting model [12] to whiten the background. Pseudo ground truth mesh registrations are obtained with an optimization-based head tracker [16].

2. Implementation Details

Table 2 provides a list of the most important hyperparameter values, and Table 1 presents the loss weights used in the different training stages. We use the AdamW optimizer [13] with an initial learning rate of $4e - 5$, a weight decay of 0.05, and a cosine learning rate scheduler that decreases the

Iteration	Stage Name	Training Datasets	Loss Weights
0 - 100k	Geometry Only	Ava-256 [15]	$w_{\text{geometry}} = 1 \times 10^{-3}$ $w_{\text{reg}} = 1 \times 10^{-3}$ $w_{\text{lpips}} = 0$ $w_{\text{L1}} = 0$ $w_{\text{SSIM}} = 0$
100k - 400k	Geometry & Apparance	Ava-256 [15]	$w_{\text{geometry}} = 1 \times 10^{-3}$ $w_{\text{reg}} = 1 \times 10^{-3}$ $w_{\text{L1}} = 0.8$ $w_{\text{SSIM}} = 0.2$ $w_{\text{LPIPS}} = 0$ until 150k, then linearly increasing to 1.0 until 200k
400k - 860k	Mixed Training	Ava-256 [15] & NeRSemble [8]	$w_{\text{geometry}} = (\text{Ava-256: } 1 \times 10^{-3}; \text{NeRSemble: } 0)$ $w_{\text{reg}} = 1 \times 10^{-3}$ $w_{\text{L1}} = 0.8$ $w_{\text{SSIM}} = 0.2$ $w_{\text{LPIPS}} = 1$

Table 1. MATCH loss weights in different training stages.

Parameter	Value
UV texture resolution $H_{\text{uv}} \times W_{\text{uv}}$	1024×1024
UV token patch size p_{uv}	16
Image token patch size p_{img}	8
Token dimension d	512
Number of input images V	12
Image resolution $H_{\text{img}} \times W_{\text{img}}$	640×512
Number of registration-guided attention blocks	6
Registration-guided attention image token count $k_{\mathcal{T}, \text{img}}$	100
Number of grouped attention blocks	6
Gaussian scale regularization target	5×10^{-4}
Gaussian opacity regularization target	0.7
Learnable UV token positional embedding dimension	512

Table 2. MATCH hyperparameters.

learning rate to 0 within 1M steps after a 1,000-step linear warm-up phase. For calculating the Sapiens feature maps of the input images, we assemble them into grids of 2×2 before feeding them into the feature extractor to save computation time.

3. Novel View Synthesis

3.1. Detailed Baseline Description

We compare our model against the baselines GPAvatar [2], Fastavatar [19], LAM [5], Avat3r [9], FaceLift [14], and CAP4D [18].

LAM [5] predicts Gaussian splats for each vertex of a subsampled 3DMM from a single input image using a transformer backbone. These can be directly driven using 3DMM parameters. Fastavatar [19] builds on this approach and enables the aggregation of information extracted from several input images. GPAvatar [2] follows a different approach and reconstructs 3D head avatars from one or several input images using a triplane representation that can be an-

imated with point-based expression fields.

FaceLift [14] trains a Gaussian Splatting Large Reconstruction Model (GS-LRM) [20] on synthetic data to predict pixel-aligned Gaussian splats from several input images. This GS-LRM is used to lift predictions of a diffusion model, which infers multi-view images from a single reference, into a 3D Gaussian splatting representation. In our comparison, we solely focus on the GS-LRM model of FaceLift, which receives the ground truth multi-view images as input. Avat3r [9] similarly regresses pixel-aligned Gaussian splats, yet they can be directly animated into new expressions through cross-attention to latent expression codes. Note that these latent expression codes are constructed from high-quality mesh registrations and texture re-projections that are obtained with closed-source software, making Avat3r inapplicable to datasets other than Ava-256. The Gaussian splats predicted by FaceLift and Avat3r are pixel-aligned and do not exhibit any semantic correspondence across frames or subjects.

CAP4D [18] uses a 3DMM-conditioned multi-view diffusion model to generate images with novel pose and expression, given one or several input images. In contrast to the other baselines, which infer 3D representations, it predicts 2D images that are not truly 3D-consistent. CAP4D only held out two subjects of the Ava-256 dataset for validation, only one of which intersects with the validation subjects from Avat3r and our method. As a consequence, we only perform a qualitative comparison with CAP4D on this one subject (see Figure 2).

Since different methods predict different crops of the face, we evaluate the methods on the maximum square crop that fits into the intersection of all bounding boxes, resized to a resolution of 512×512 , and mask out the torso and shoulders using a pretrained segmentation network [7].



Figure 2. Novel view synthesis comparison against CAP4D on Ava-256.

V_{TEMPEH}	V_{MATCH}	LPIPS ↓	CSIM ↑	PSNR ↑	SSIM ↑	L1 ↓	L2 ↓
12	2	0.213	0.866	20.503	0.795	0.042	0.013
12	4	0.212	0.865	21.078	0.798	0.039	0.012
12	8	0.195	0.907	22.350	0.816	0.034	0.010
12	12	0.187	0.918	23.032	0.825	0.032	0.009
12	16	0.182	0.923	23.367	0.831	0.032	0.009
2	2	0.261	0.689	15.996	0.731	0.071	0.033
4	4	0.230	0.798	19.289	0.775	0.047	0.017
8	8	0.198	0.903	22.076	0.812	0.035	0.010
12	12	0.187	0.918	23.032	0.825	0.032	0.009
16	16	0.182	0.924	23.265	0.831	0.032	0.009

Table 3. Quantitative ablation of the number of input views to MATCH evaluated on Ava-256. We evaluate two scenarios: *i*) Changing the number of input views to MATCH while keeping the number of inputs to the coarse mesh registration model (TEMPEH) at the default ($V = 12$). *ii*) Changing the number of input views for both TEMPEH and MATCH.

3.2. Further qualitative comparisons

Figure 2 presents the qualitative comparison with CAP4D on the one intersecting validation subject. We observe that our method predicts reconstructions with better identity preservation and expression fidelity. Figure 12 and Figure 13 provide additional comparisons with the remaining baselines on samples from Ava-256 and NeRSemble respectively. As discussed in the main paper, our method exhibits superior synthesis quality.



Figure 3. Additional semantic editing results. From top to bottom: Transferring beard and lips, eyes, and hairstyle.

3.3. Additional Ablations

Number of input images. Figure 4 qualitatively evaluates the impact of the number of input views on the synthesis result. We conduct two lines of experiments: *i*) Keeping the number of input images to the coarse mesh registration model (TEMPEH) at the default ($V = 12$) while only changing the number of input views to MATCH. This evaluates the actual impact of the number of input views on our method in isolation. *ii*) TEMPEH and MATCH receive the same number of input images. This is the more realistic scenario, but entangles the sensitivity of TEMPEH to few input images with MATCH's.

We find that MATCH is highly robust to few input images and can generate plausible reconstructions even for two input images, assuming high-quality geometry initialization. However, TEMPEH's geometry prediction degrades significantly for two views, resulting in a low-quality reconstruction for the combined scenario. If TEMPEH and MATCH receive the same number of input views, starting from four images, plausible results are produced. Fine details improve as more input views are added. This is confirmed by Table 3 and Figure 6 (top), which show improving LPIPS scores as the number of views increases. Fig-



Figure 4. Qualitative ablation study for the number of input views to MATCH. We evaluate two scenarios. Top: Changing the number of input views to MATCH while keeping the number of inputs to the coarse mesh registration model (TEMPEH) at the default ($V = 12$). Bottom: Changing the number of input views for both TEMPEH and MATCH.



Figure 5. Qualitative ablation study for $k_{\mathcal{T},\text{img}}$, i.e., the number of image tokens that each UV token attends to in the registration-guided attention blocks. The default value is $k_{\mathcal{T},\text{img}} = 100$.

ure 4 (bottom) reports the inference speed. As discussed in the main paper, while the computational complexity scales quadratically with the number of input images for dense attention between all UV and image tokens, our method’s complexity increases only linearly. Especially for high numbers of input images, this results in a considerable improvement of inference speed ($1.8\times$ acceleration compared to dense attention for 16 input images). We found 12 input images to be a good compromise between inference speed and synthesis quality, running at a framerate of 2 fps.

Registration-guided attention context length. We ablate the effect of $k_{\mathcal{T},\text{img}}$, i.e., the number of image tokens that each UV token attends to in the registration-guided

attention blocks. The quantitative comparison in Table 4 shows minor improvements as we decrease $k_{\mathcal{T},\text{img}}$. However, we did not observe pronounced qualitative differences as shown in Figure 5.

Robustness to coarse mesh errors. MATCH uses a coarse mesh estimated by TEMPEH as initialization. In practice, TEMPEH exhibits moderate inaccuracies, which MATCH can recover from, see Figure 7 (left), producing plausible results. When perturbing the coarse mesh by constant vertex offsets Δx , Figure 7 (right), artifacts only appear for $\Delta x \geq 20\text{mm}$, which is $7\times$ higher than the average point-to-surface distance of TEMPEH.

Image token self attention. MATCH performs self-

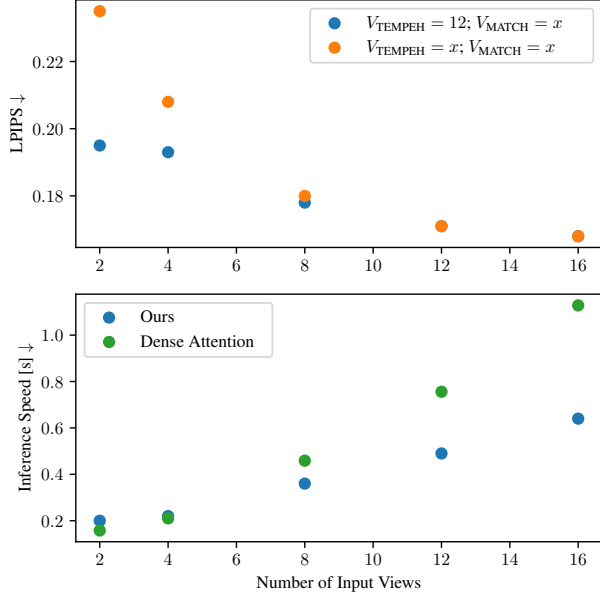


Figure 6. Top: Quantitative ablation study for the number of input views to MATCH on Ava-256. We evaluate two scenarios: *i*) Changing the number of input views to MATCH while keeping the number of inputs to the coarse mesh registration model (TEMPEH) at the default ($V = 12$). *ii*) Changing the number of input views for both TEMPEH and MATCH. Bottom: Inference speed comparison between our model with the novel registration-guided attention versus a version with dense attention across all UV and image tokens.

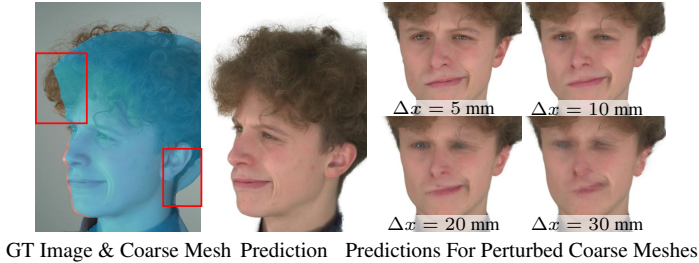


Figure 7. Robustness to errors in the coarse TEMPEH mesh.

attention between the image tokens of each image to enable image-space signal processing inside the grouped attention blocks. We found that this self-attention can be skipped without harming the model performance while increasing the inference speed by 8%.

4. In-the-wild application.

While MATCH was trained on calibrated studio-captures with uniform lighting and known camera parameters, we found that it generalizes to in-the-wild captures and yields high-quality reconstructions, see Figure 8 (top). The input

$k_{\mathcal{T}, \text{img}}$	LPIPS ↓	CSIM ↑	PSNR ↑	SSIM ↑	L1 ↓	L2 ↓
25	0.184	0.926	22.908	0.830	0.032	0.009
50	0.183	0.924	23.164	0.830	0.031	0.008
100	0.187	0.918	23.032	0.825	0.032	0.009
150	0.185	0.919	22.951	0.826	0.032	0.009

Table 4. Quantitative ablation study for $k_{\mathcal{T}, \text{img}}$, i.e., the number of image tokens that each UV token attends to in the registration-guided attention blocks. The evaluations were performed on the Ava-256 dataset. The default value is $k_{\mathcal{T}, \text{img}} = 100$.

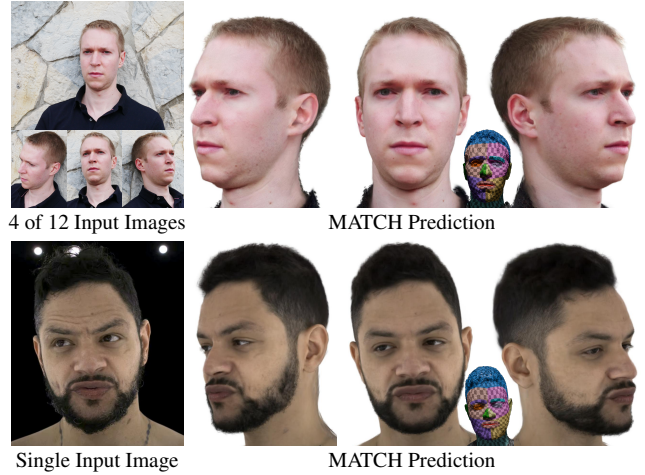


Figure 8. In-the-wild (top) and single-image (bottom) results.

images were captured with an off-the-shelf camera in an outdoor environment, and we used COLMAP [4] to estimate the camera parameters.

5. Single-image inference.

MATCH is not trained to hallucinate unobserved regions and shows artifacts for two input images only (see Figure 4). However, we can follow FaceLift [14] to generate additional views from a single input image with the 2D prior of CAP4D and input these to MATCH. Figure 8 (bottom) shows that this yields high-quality reconstructions.

6. Additional Results for Interpolation, Editing, and Expression Transfer

Figure 14, Figure 3, Figure 9 present further results for interpolation, semantic editing, and expression transfer, respectively. We observe smooth interpolations between samples and plausible editing results for swapping beard, eyes, and hairstyle. As discussed in the main paper, the arithmetic expression transfer approach, where the residual of Gaussian maps for an expressive and a neutral frame of a target subject is added to the neutral reconstruction of a source identity, can result in uncanny results for extreme expressions and dissimilar identities. A less simplistic method,

Corresp. Dist. [mm] ↓	Full Head	Face	Ears	Eyes	Mouth	Scalp
TEMPEH [1] / Ours	8.9 / 8.0	5.4 / 4.8	10.8 / 9.1	2.5 / 2.1	2.8 / 2.7	13.5 / 12.5

Table 5. Quantitative correspondence evaluation.



Figure 9. Additional expression transfer results. Note that we only aim to transfer the oral expression and do not apply any modifications to other regions, e.g., the eyes.

e.g., a conditional VAE [15], would be a more suitable choice for this challenging task.

7. Quantitative Correspondence Evaluation.

We quantify the semantic correspondence of MATCH’s predictions using Ava-256’s ground truth mesh registrations. Table 5 reports the Euclidean distance between the center of each predicted Gaussian and its corresponding target location obtained through barycentric interpolation on 1,000 samples. The same interpolation is done to evaluate TEMPEH’s results. We find that MATCH produces superior correspondence.

8. Additional Material for the Subject-Specific Avatars

8.1. Detailed Avatar Creation Procedure

This section illustrates the changes applied to GEM’s [21] procedure to create a lightweight animatable avatar from a set of Gaussian splat textures predicted by MATCH. Ablations

on the effect of the individual changes are presented in Figure 11 and Table 7, which are discussed in Section 8.5. Figure 10 provides an overview of the resulting procedure.

i) Skip Tracking & CNN-based Avatar Training: Since MATCH directly predicts Gaussian splats that are in correspondence across frames, we can skip the time-expensive procedure of tracking and CNN-based head avatar training, which drastically reduces the time to create a lightweight head avatar (see Table 6). Since the reconstruction of the PCA basis requires unposed Gaussians in a canonical space, we have to unpose MATCH’s predictions. To this end, we extract the Ava-256 mesh by sampling the texture of predicted 3D Gaussian locations at the template vertices’ UV coordinates. We then convert the Ava-256 mesh to the topology of FLAME [11], a publicly available 3D morphable model (3DMM), using a fixed mapping of vertex locations. The 3DMM parameters are obtained by optimizing FLAME’s vertices against our vertex predictions using a Huber loss [6]. Finally, we can use the obtained FLAME pose parameters to apply inverse linear blend skinning to transform the Gaussians predicted by MATCH into an unposed canonical space on which we perform the PCA decomposition. Note that during this unposing operation, the jaw articulation is neutralized as well. For this reason, we train GEM’s expression encoder to also predict the jaw pose in addition to the Eigen-coefficients. To reduce the compute cost and memory requirements of the PCA decomposition, we use a version of MATCH that predicts Gaussian textures with a reduced resolution of 512×512 .

ii) Modality-agnostic PCA: GEM creates separate PCAs for each of the Gaussian’s modalities (rotation, position, opacity, and scale). However, we found that this formulation misses crucial correlations between the modalities (e.g., raised eyebrows should correlate with color changes in wrinkles on the forehead). This is resolved by modelling all Gaussian modalities in a joint PCA.

iii) Enable dynamic colors: GEM disables dynamic color changes to promote semantic correspondence of Gaussians across frames. For MATCH, however, this is neither feasible nor practical, since it must predict dynamically changing colors to reconstruct the appearance of different subjects, and intrinsically exhibits high semantic correspondence across subjects and frames. As such, we drop the constraint of static colors during the PCA reconstruction and refinement. The only exception to this is the interior of the mouth cavity. Since the ground truth mesh registrations on the Ava-256 dataset simplify the oral cavity as planar surfaces between the lips, the semantic correspondence of MATCH’s predictions in this area is limited. Plausible reconstructions are achieved through intricately changing colors, opacities, and scales. We found that naïvely using the lightweight expression MLP to predict these complex dynamics yields test-time artifacts. To alleviate this problem,

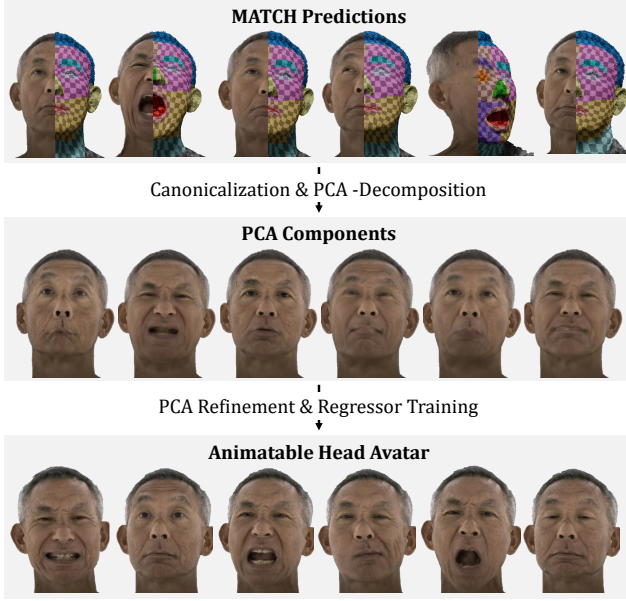


Figure 10. Procedure to create subject-specific head avatars from a sequence of Gaussian splat textures predicted by MATCH.

we fix the colors, scales, and opacities of the Gaussians in the oral cavity to their mean calculated across all training frames.

iv) *Mean Refinement*: GEM only refines the PCA basis vectors \mathbf{B}_i against the target images using photometric losses. We found it beneficial to also refine the PCA means μ_i during that stage.

8.2. Detailed Baseline Description

We compare our avatars with the optimization-based methods GaussianAvatars [17], RGBAvatar [10], and GEM [21]. GaussianAvatars optimizes Gaussian splats that are rigged to a parametric morphable face model against multi-view videos. RGBAvatar follows a similar approach, yet it also estimates Gaussian blendshapes from the face model parameters that can model dynamic appearance and geometry changes beyond the underlying face model. GEM first optimizes a CNN-based high-quality head avatar, which is then distilled into a lightweight, blendshape-based representation that can be directly animated from driving images.

GaussianAvatars and RGBAvatar can be directly driven with parameters of the FLAME 3DMM. For image-based animation, we estimate these parameters with EMOCA [3], a state-of-the-art 3DMM estimation method, which is also used as a pretrained feature extractor to drive GEM and our method. We found it beneficial for RGBAvatar to also use the EMOCA predictions during training. The performances for self- and cross-reenactment are evaluated on five subjects from the Ava-256 dataset. All methods are trained on

a subset of the available sequences, avoiding extreme head and shoulder movements, protruded tongues, and isolated eye movements with a neutral face, while leaving out the `EXP_free_face` sequence for validation. Since we only aim to extract facial expressions from the driving image, not the global rigid transformation, we use the ground truth global pose from the VHAP tracking for the baselines and from MATCH’s registrations for our method during evaluation.

8.3. Detailed Reconstruction Time Analysis

Table 6 presents a detailed breakdown of the time cost distribution across the individual stages of head avatar reconstruction for each method. The measurements were taken on a representative training sequence with 3212 frames using a compute node with a single NVIDIA A100 40GB GPU, 16 CPUs, and 500GB of RAM. To ensure full GPU usage during VHAP tracking, we ran two processes in parallel. File system operations, e.g., data loading and writing, were excluded from all timing computations since they are highly system-dependent. We find that the major bottleneck of the baseline’s reconstruction time, especially for RGBAvatar, lies in the multi-view head tracking. While RGBAvatar reports an impressive reconstruction time of only 80s for the monocular setting, in the multi-view setting, they require optimization-based tracking with VHAP [16], which takes 10.65h on a representative 3212 frame training sequence, while the avatar optimization time increases to 0.75h¹. GEM’s multi-stage approach of first tracking a parametric head model, then optimizing a high-quality head avatar, followed by a distillation, even increases the total reconstruction time per avatar to 45.3h in our setup. Instead, MATCH allows for skipping the lengthy optimization-based mesh registration by directly predicting registered Gaussians from the multi-view images, which takes 0.53h for the entire training sequence compared to 10.65h of optimization-based tracking with VHAP. Unposing and PCA decomposition take 0.16 hours, such that we can start the refinement of the blendshapes and training of the expression regressor even before any of the baselines has completed registering just one 10th of the training frames.

8.4. Further qualitative comparisons

Figure 15 and Figure 16 present further results of the personalized head avatars for self- and cross-reenactment respectively.

8.5. Ablation Study

Figure 11 and Table 7 present qualitative and quantitative ablation studies of the changes applied to GEM [21] to

¹Experiments conducted with hyperparameters from the official code base.

Method	GA [17]	RGBAvatar [10]	GEM [21]	Ours
Stage-Wise Durations	VHAP Tracking: 10.65h Avatar Optimization: 4.83h	VHAP Tracking: 10.65h Avatar Optimization: 0.75h	VHAP Tracking: 10.65h CNN-Avatar Optimization: 27.70h Regressor Training: 6.94h	Coarse Mesh Registration: 0.09h MATCH Inference: 0.44h Canonicalization & PCA Decomp.: 0.16h Emoca & Deca Inference: 0.05h PCA Refinement: 2.75h Expression Regressor Training: 1.14h
Total Reconstruction Time	15.48h	11.40h	45.29h	4.63 h

Table 6. Head avatar reconstruction time breakdown. The measurements were conducted on a representative training sequence with 3,212 frames.

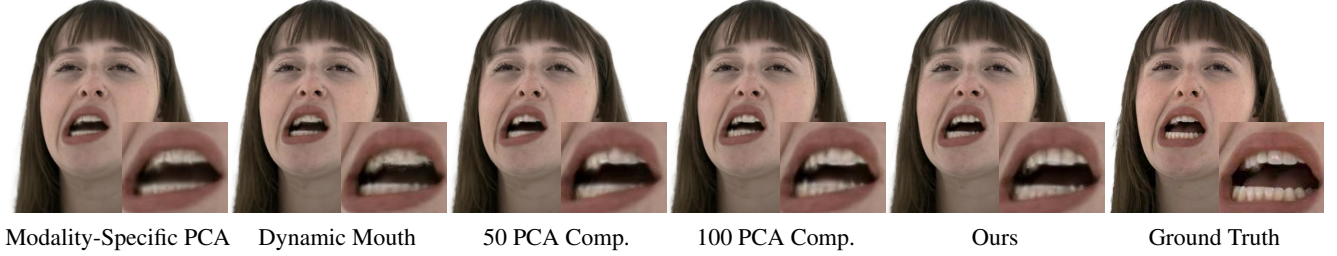


Figure 11. Qualitative ablation study of the changes applied to GEM [21] to create subject-specific head avatars from MATCH’s predictions. Ours uses 150 PCA components.

	Self-Reenactment					Cross-Reenactment	
	LPIPS ↓	CSIM ↑	SSIM ↑	L1 ↓	PSNR ↑	CSIM ↑	EmoL1 ↓
Modality-Specific PCAs	0.180	0.879	0.816	0.027	24.376	0.815	9.849
Dynamic Mouth	0.174	0.878	0.809	0.027	24.112	0.811	9.792
# PCA Comp. = 50	0.177	0.876	0.808	0.027	24.126	0.814	9.891
# PCA Comp. = 100	0.175	0.880	0.809	0.027	24.112	0.814	9.907
Ours	0.174	0.880	0.809	0.027	24.122	0.813	9.837

Table 7. Quantitative ablation study of the changes applied to GEM [21] to create subject-specific head avatars from MATCH’s predictions. By default, we use 150 PCA components.

create subject-specific head avatars from MATCH’s predictions. We find that employing separate PCAs for the individual Gaussian modalities (‘Modality-Specific PCAs’), i.e., location, color, scale, rotation, and opacity, yields inferior results compared to jointly modeling all attributes in a single PCA. This aligns with the intuition that the different Gaussian attributes are highly correlated (e.g., raising the eyebrows results in darker colors for wrinkles on the forehead). Modeling the mouth interior with Gaussians with dynamically changing color, opacity, and scale (‘Dynamic Mouth’) does not change the quantitative scores significantly, yet slightly reduces the faithfulness of extreme expressions at test time (see Figure 11). We deduce that the lightweight image-based expression encoder fails to learn the intricate dynamics of the highly dynamic mouth interior Gaussians and benefits from additional consistency constraints enforced through static colors, opacities, and scales in this region. Increasing the number of PCA components improves the perceptual quality by adding high-frequency details. Note that even with the highest number of PCA components that we test, i.e., our default value of 150, we

still use fewer components than GEM’s modality-specific PCAs with a total of 180 components.

9. Ethical Considerations

Our method relies on multi-view studio captures with calibrated cameras, ensuring that all participants were aware of and consented to data collection. However, with the emergence of generative multi-view models such as CAP4D [18], similar data could be fabricated synthetically. This raises potential ethical concerns regarding consent and misuse, which we strongly discourage.



Figure 12. Additional novel view synthesis results on Ava-256 [15].



GPAvatar [2] FastAvatar [19] LAM [5] FaceLift [14] Ours (Ava) Ours (NeRSemle) Ours Ground Truth

Figure 13. Additional novel view synthesis results on NeRSemle [8]. Ours (Ava) / Ours (NeRSemle) are trained on Ava-256 [15] and NeRSemle only, respectively.



Figure 14. Additional interpolation results. γ denotes the interpolation factor.



Figure 15. Additional self-reenactment results for the personalized head avatars.



Figure 16. Additional cross-reenactment results for the personalized head avatars.

References

- [1] Timo Bolkart, Tianye Li, and Michael J. Black. Instant multi-view head capture through learnable registration. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 768–779, 2023. 6
- [2] Xuangeng Chu, Yu Li, Ailing Zeng, Tianyu Yang, Lijian Lin, Yunfei Liu, and Tatsuya Harada. GPAvatar: Generalizable and precise head avatar from image(s). In *The Twelfth International Conference on Learning Representations*, 2024. 2, 9, 10
- [3] Radek Daněček, Michael J Black, and Timo Bolkart. EMOCA: Emotion driven monocular face capture and animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20311–20322, 2022. 7
- [4] Alex Fisher, Ricardo Cannizzaro, Madeleine Cochrane, Chatura Nagahawatte, and Jennifer L Palmer. COLMAP: A memory-efficient occupancy grid mapping framework. *Robotics and Autonomous Systems*, 142:103755, 2021. 5
- [5] Yisheng He, Xiaodong Gu, Xiaodan Ye, Chao Xu, Zhengyi Zhao, Yuan Dong, Weihao Yuan, Zilong Dong, and Liefeng Bo. Lam: Large avatar model for one-shot animatable gaussian head. In *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers*, pages 1–13, 2025. 2, 9, 10
- [6] Peter J Huber. Robust estimation of a location parameter. In *Breakthroughs in statistics: Methodology and distribution*, pages 492–518. Springer, 1992. 6
- [7] Rawal Khirodkar, Timur Bagautdinov, Julieta Martinez, Su Zhaoen, Austin James, Peter Selednik, Stuart Anderson, and Shunsuke Saito. Sapiens: Foundation for human vision models. In *European Conference on Computer Vision*, pages 206–228. Springer, 2024. 2
- [8] Tobias Kirschstein, Shenhan Qian, Simon Giebenhain, Tim Walter, and Matthias Nießner. Nersemble: Multi-view radiance field reconstruction of human heads. *ACM Trans. Graph.*, 42(4), 2023. 1, 2, 10
- [9] Tobias Kirschstein, Javier Romero, Artem Sevastopolsky, Matthias Nießner, and Shunsuke Saito. Avat3r: Large animatable gaussian reconstruction model for high-fidelity 3d head avatars. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 12089–12100, 2025. 1, 2, 9
- [10] Linzhou Li, Yumeng Li, Yanlin Weng, Youyi Zheng, and Kun Zhou. Rgbavatar: Reduced gaussian blendshapes for online modeling of head avatars. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 10747–10757, 2025. 7, 8, 12, 13
- [11] Tianye Li, Timo Bolkart, Michael J. Black, Hao Li, and Javier Romero. Learning a model of facial shape and expression from 4D scans. *Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 36(6):194:1–194:17, 2017. 6
- [12] Shanchuan Lin, Andrey Ryabtsev, Soumyadip Sengupta, Brian Curless, Steve Seitz, and Ira Kemelmacher-Shlizerman. Real-time high-resolution background matting. *arXiv*, pages arXiv–2012, 2020. 1
- [13] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 1
- [14] Weijie Lyu, Yi Zhou, Ming-Hsuan Yang, and Zhixin Shu. Facelift: Learning generalizable single image 3d face reconstruction from synthetic heads. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12691–12701, 2025. 2, 5, 9, 10
- [15] Julieta Martinez, Emily Kim, Javier Romero, Timur Bagautdinov, Shunsuke Saito, Shou-I Yu, Stuart Anderson, Michael Zollhöfer, Te-Li Wang, Shaojie Bai, Chenghui Li, Shih-En Wei, Rohan Joshi, Wyatt Borsos, Tomas Simon, Jason Saragih, Paul Theodosis, Alexander Greene, Anjani Josyula, Silvio Mano Maeta, Andrew I. Jewett, Simon Venstain, Christopher Heilman, Yueh-Tung Chen, Sidi Fu, Mohamed Ezzeldin A. Elshaer, Tingfang Du, Longhua Wu, Shen-Chi Chen, Kai Kang, Michael Wu, Youssef Emad, Steven Longay, Ashley Brewer, Hitesh Shah, James Booth, Taylor Koska, Kayla Haidle, Matt Andromalos, Joanna Hsu, Thomas Dauer, Peter Selednik, Tim Godisart, Scott Ardisson, Matthew Cipperly, Ben Humberston, Lon Farr, Bob Hansen, Peihong Guo, Dave Braun, Steven Krenn, He Wen, Lucas Evans, Natalia Fadeeva, Matthew Stewart, Gabriel Schwartz, Divam Gupta, Gyeongsik Moon, Kaiwen Guo, Yuan Dong, Yichen Xu, Takaaki Shiratori, Fabian Prada, Bernardo R. Pires, Bo Peng, Julia Buffalini, Autumn Trimble, Kevyn McPhail, Melissa Schoeller, and Yaser Sheikh. Codec Avatar Studio: Paired Human Captures for Complete, Driveable, and Generalizable Avatars. *NeurIPS Track on Datasets and Benchmarks*, 2024. 1, 2, 6, 9, 10
- [16] Shenhan Qian. Vhap: Versatile head alignment with adaptive appearance priors, 2024. 1, 7
- [17] Shenhan Qian, Tobias Kirschstein, Liam Schoneveld, Davide Davoli, Simon Giebenhain, and Matthias Nießner. GaussianAvatars: Photorealistic head avatars with rigged 3D gaussians. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20299–20309, 2024. 7, 8, 12, 13
- [18] Felix Taubner, Ruihang Zhang, Mathieu Tuli, and David B. Lindell. CAP4D: Creating animatable 4D portrait avatars with morphable multi-view diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5318–5330, 2025. 2, 3, 8
- [19] Yue Wu, Yufan Wu, Wen Li, Yuxi Lu, Kairui Feng, and Xuanhong Chen. Fastavatar: Towards unified fast high-fidelity 3d avatar reconstruction with large gaussian reconstruction transformers. *arXiv preprint arXiv:2508.19754*, 2025. 2, 9, 10
- [20] Kai Zhang, Sai Bi, Hao Tan, Yuanbo Xiangli, Nanxuan Zhao, Kalyan Sunkavalli, and Zexiang Xu. Gs-lrm: Large reconstruction model for 3D Gaussian splatting. In *European Conference on Computer Vision*, pages 1–19. Springer, 2024. 2
- [21] Wojciech Zielonka, Timo Bolkart, Thabo Beeler, and Justus Thies. Gaussian eigen models for human heads. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 15930–15940, 2025. 6, 7, 8, 12, 13