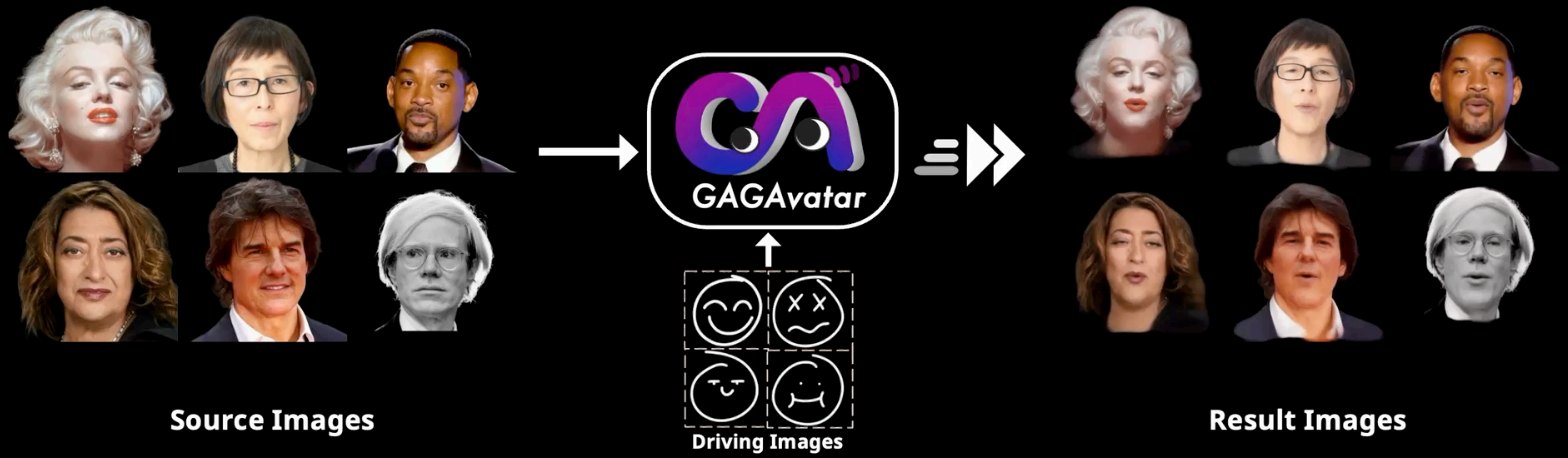
Generalizable and Animatable Gaussian Head Avatar

Xuangeng Chu[1], Tatsuya Harada[1,2]
The University of Tokyo[1], RIKEN AIP[2]

UTokyo

RIKEN

Overview



Key Ideas

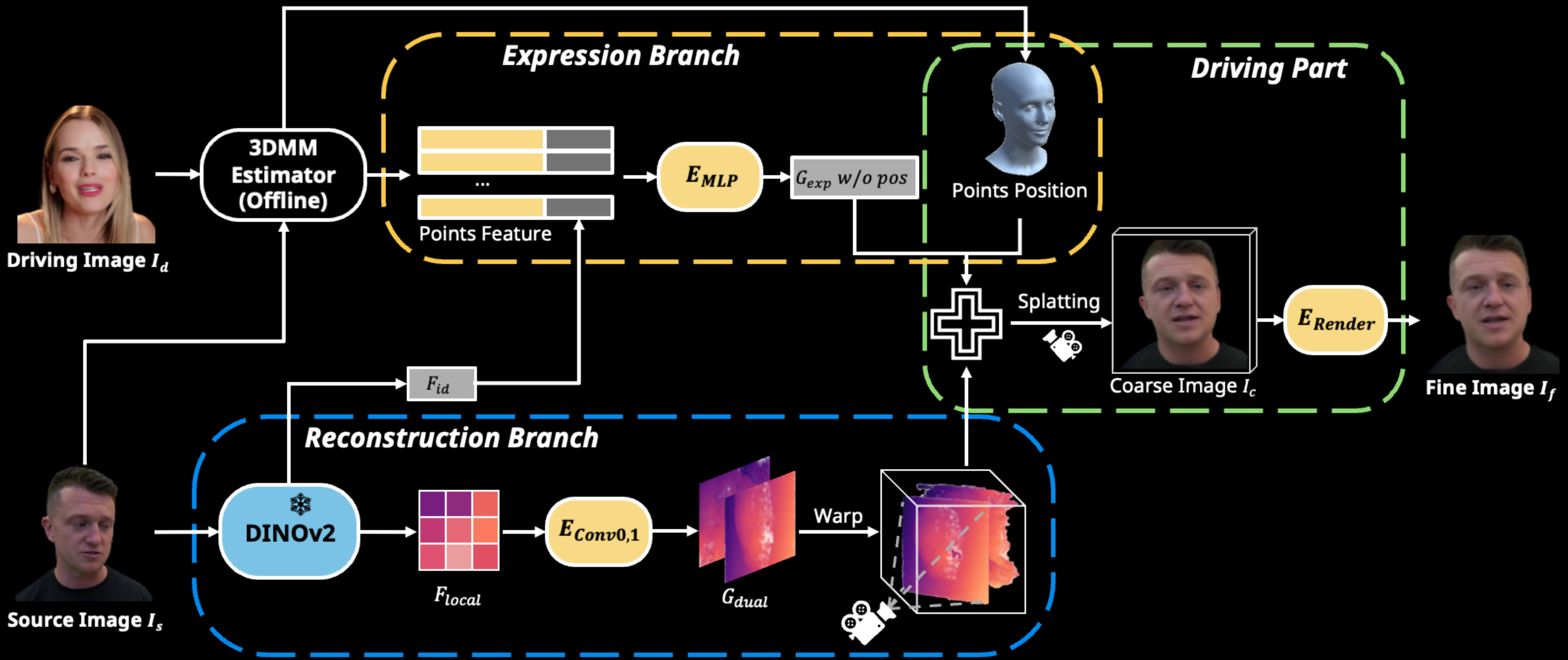
Our goal is to build a head avatar framework that achieves **single forward reconstruction** with **one image** and **real-time reenactment**.

- To achieve this, we propose a dual lifting method that lifts 3DGS from a single image.
- Then we blends 3DMM-based expression Gaussians to achieve re-reenactment.
- We also use 3DMM prior to constraint the lifting process and using a fast neural rendering module to refine the Gaussian Splatting result.

Different from Related Works

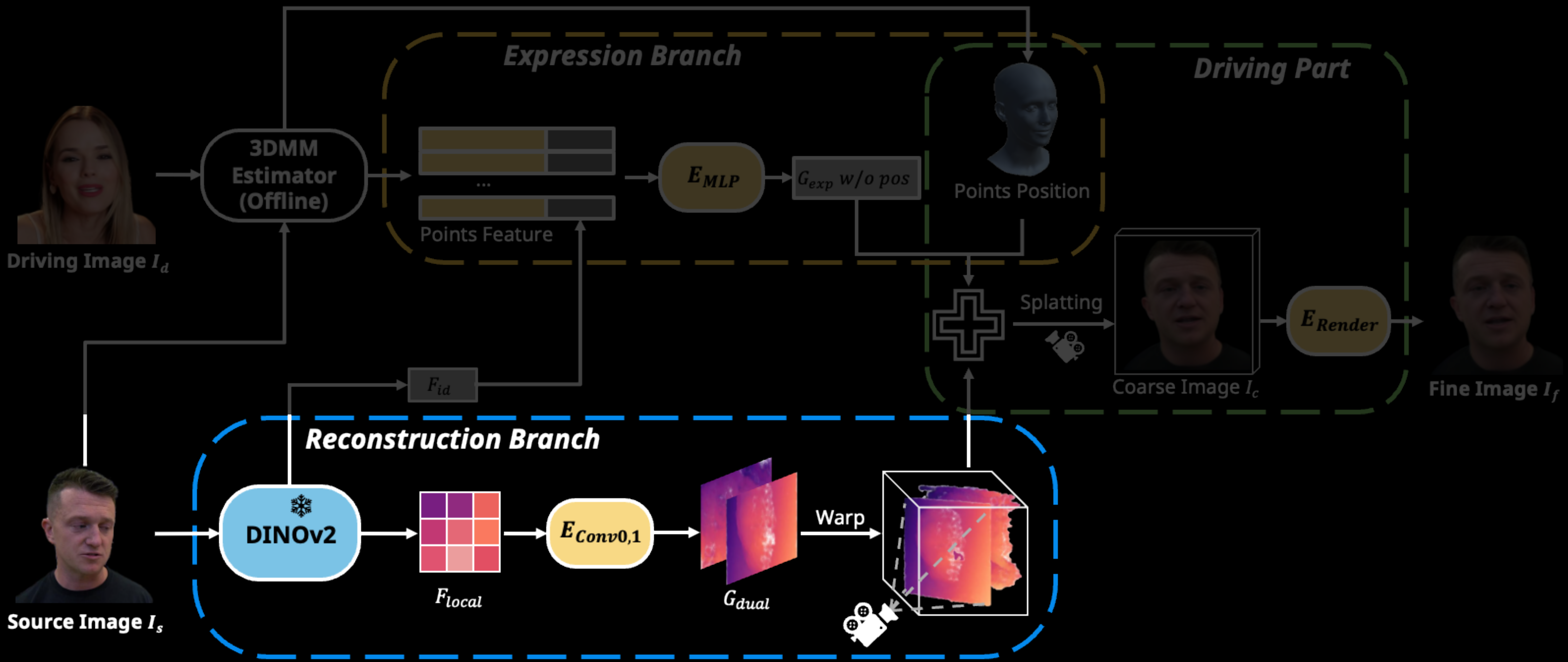
	One-shot reconstruction	No per-ID optimization	Real-time reenactment
ROME	✔	✔	✘
OTAvatar	✔	✘	✘
HideNeRF	✔	✔	✘
GOHA	✔	✔	✘
GPAvatar	✔	✔	✘
Real3DPortrait	✔	✔	✘
Portrait4D-v2	✔	✔	✘
Gaussian Head Avatar	✘	✘	✔
FlashAvatar	✘	✘	✔
GAGAvatar (Ours)	✔	✔	✔

Method

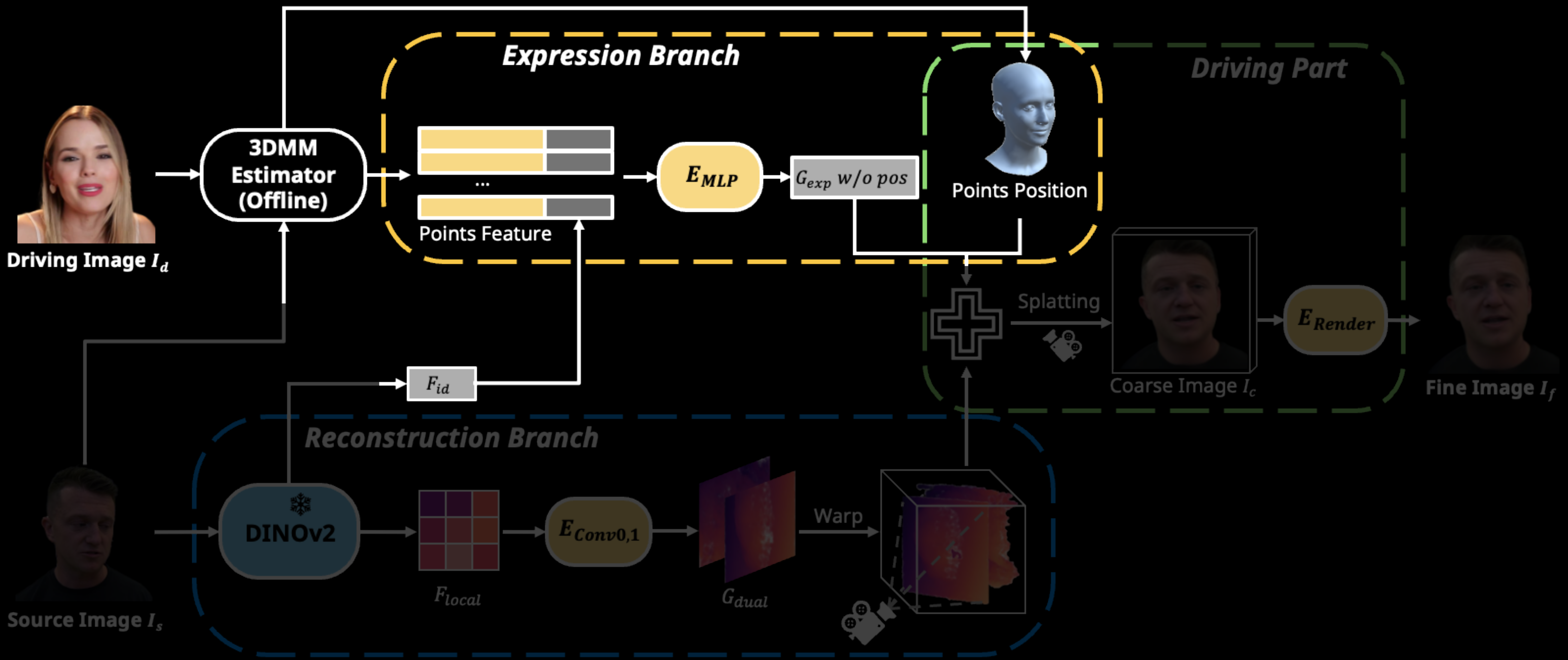


Expression Branch
Driving Part
3DMM Estimator (Offline)
Driving Image I_d
...
Points Feature
E_{MLP}
G_{exp} w/o pos
Points Position
Splatting
Coarse Image I_c
E_{Render}
Fine Image I_f
F_{id}
Reconstruction Branch
DINOv2
F_{local}
$E_{Conv0,1}$
G_{dual}
Warp
Source Image I_s

Method

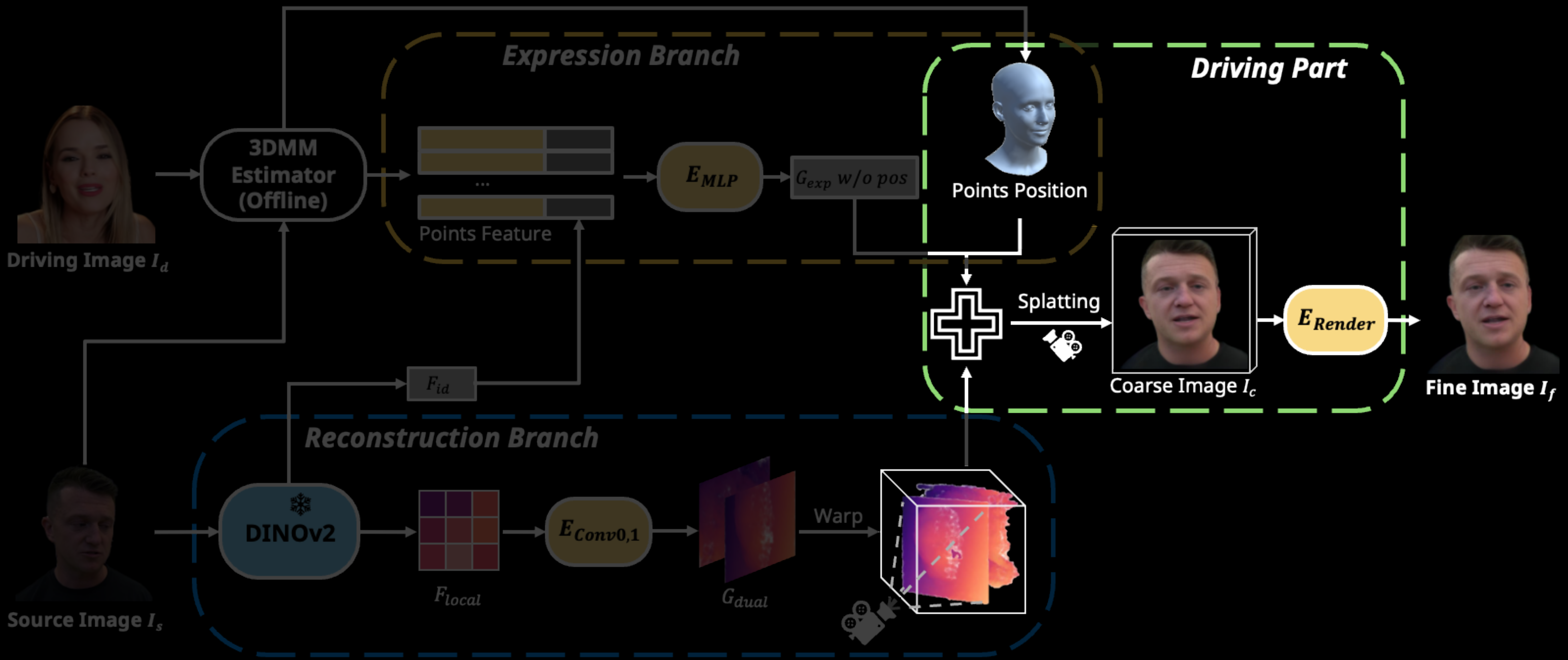

Expression Branch
Driving Part
3DMM Estimator (Offline)
Driving Image I_d
Points Feature
E_{MLP}
G_{exp} w/o pos
Points Position
Splatting
Coarse Image I_c
E_{Render}
Fine Image I_f
F_{id}
Reconstruction Branch
Source Image I_s
DINOv2
F_{local}
$E_{Conv0,1}$
G_{dual}
Warp

Method



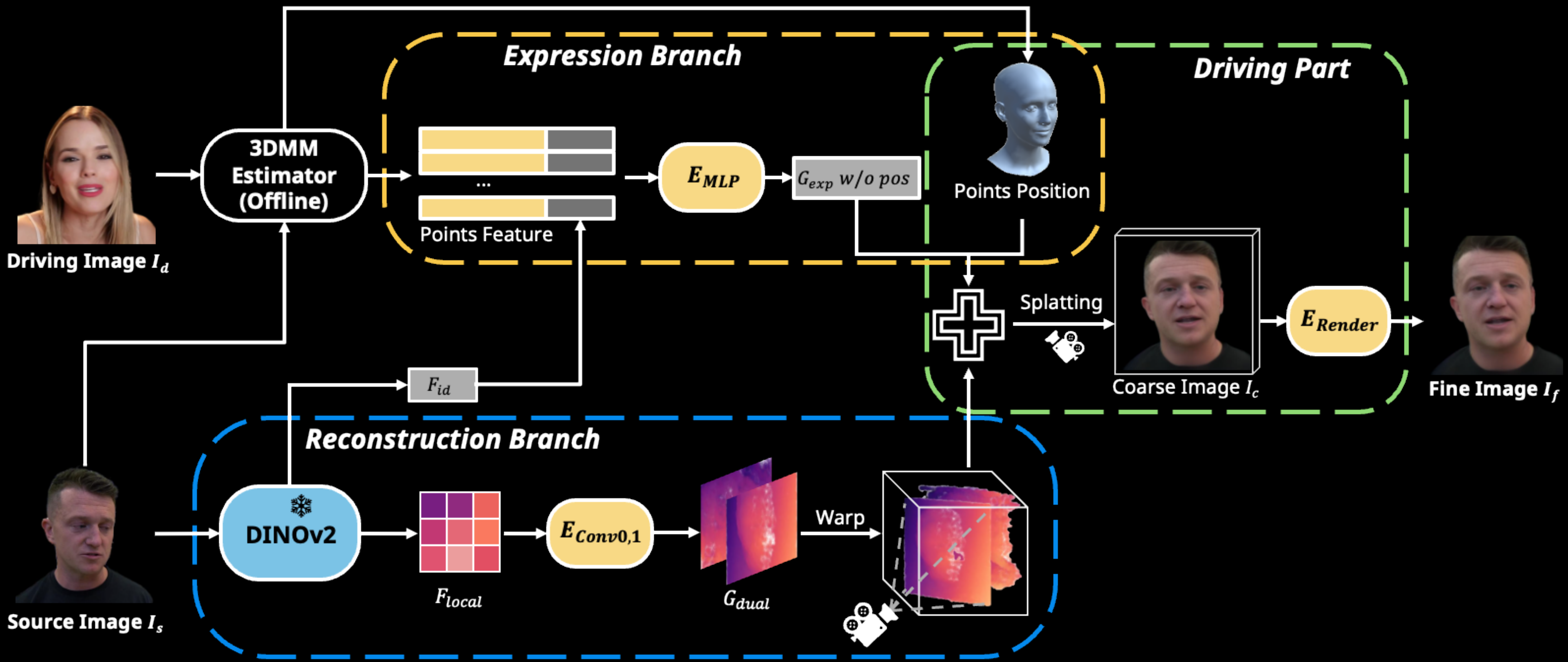
Expression Branch
Driving Part
Driving Image I_d
3DMM Estimator (Offline)
...
Points Feature
E_{MLP}
G_{exp} w/o pos
Points Position
Splatting
Coarse Image I_c
E_{Render}
Fine Image I_f
F_{id}
Reconstruction Branch
DINOv2
F_{local}
$E_{Conv0,1}$
G_{dual}
Warp
Source Image I_s

Method



Expression Branch
Driving Part
3DMM Estimator (Offline)
Points Feature
E_{MLP}
G_{exp} w/o pos
Points Position
Driving Image I_d
Splatting
Coarse Image I_c
E_{Render}
Fine Image I_f
F_{id}
Reconstruction Branch
DINOv2
F_{local}
$E_{Conv0,1}$
G_{dual}
Warp
Source Image I_s

Method



Expression Branch
Driving Part
3DMM Estimator (Offline)
...
Points Feature
E_{MLP}
G_{exp} w/o pos
Points Position
Driving Image I_d
Splatting
E_{Render}
Coarse Image I_c
Fine Image I_f
F_{id}
Reconstruction Branch
DINOv2
F_{local}
$E_{Conv0,1}$
G_{dual}
Warp
Source Image I_s

Data & Training

- Data
  - We use **video** data from **VFHQ** to train our model.

- Training process
  - All frames are tracked with head tracker to get FLAME params and camera pose.
  - During training, we sample two frames from the same video, one as the source image and the other as the driving image and target image.
  - Training target
    - $\mathscr{L}_{image} = ||I_c - I_t|| + ||I_f - I_t|| + \lambda_p(||\varphi(I_c) - \varphi(I_t)|| + ||\varphi(I_f) - \varphi(I_t)||)$
      - We require the prediction image to be consistent with the target image.
    - $\mathscr{L}_{lifting} = ||P_{3dmm} - \left\{ argmin_{q \in G_{pos}} \|p - q\| \mid p \in P_{3dmm} \right\}||$
      - We require the lifting point to be close to the 3DMM vertices.

Results

Our method works well in reconstruction quality and expression accuracy while achieving real-time rendering speed.

Table 1: Quantitative results on the VFHQ [Xie et al., 2022] dataset. We use colors to denote the first, second and third places respectively.

Method	Self Reenactment							Cross Reenactment		
	PSNR↑	SSIM↑	LPIPS↓	CSIM↑	AED↓	APD↓	AKD↓	CSIM↑	AED↓	APD↓
StyleHeat [Yin et al., 2022]	19.95	0.726	0.211	0.537	0.199	0.385	7.659	0.407	0.279	0.551
ROME [Khakhulin et al., 2022]	19.96	0.786	0.192	0.701	0.138	0.186	4.986	0.530	0.259	0.277
OTAvatar [Ma et al., 2023]	17.65	0.563	0.294	0.465	0.234	0.545	18.19	0.364	0.324	0.678
HideNeRF [Li et al., 2023a]	19.79	0.768	0.180	0.787	0.143	0.361	7.254	0.514	0.277	0.527
GOHA [Li et al., 2023b]	20.15	0.770	0.149	0.664	0.176	0.173	6.272	0.518	0.274	0.261
CVTHead [Ma et al., 2024]	18.43	0.706	0.317	0.504	0.186	0.224	5.678	0.374	0.261	0.311
GPAvatar [Chu et al., 2024]	21.04	0.807	0.150	0.772	0.132	0.189	4.226	0.564	0.255	0.328
Real3DPortrait [Ye et al., 2024]	20.88	0.780	0.154	0.801	0.150	0.268	5.971	0.663	0.296	0.411
Portrait4D [Deng et al., 2024a]	20.35	0.741	0.191	0.765	0.144	0.205	4.854	0.596	0.286	0.258
Portrait4D-v2 [Deng et al., 2024b]	21.34	0.791	0.144	0.803	0.117	0.187	3.749	0.656	0.268	0.273
Ours	21.83	0.818	0.122	0.816	0.111	0.135	3.349	0.633	0.253	0.247

Table 3: Ablation results on the VFHQ [Xie et al., 2022] dataset.

Method	Self Reenactment							Cross Reenactment		
	PSNR↑	SSIM↑	LPIPS↓	CSIM↑	AED↓	APD↓	AKD↓	CSIM↑	AED↓	APD↓
one-plane lifting	21.34	0.802	0.158	0.781	0.127	0.170	3.810	0.581	0.272	0.290
w/o F_{id}	21.13	0.807	0.155	0.774	0.125	0.155	3.722	0.537	0.270	0.272
w/o neural renderer	20.34	0.789	0.138	0.788	0.147	0.202	4.763	0.623	0.300	0.353
w/o $\mathcal{L}_{lifting}$	21.64	0.812	0.148	0.800	0.119	0.151	3.563	0.620	0.261	0.252
Ours	21.83	0.818	0.122	0.816	0.111	0.135	3.349	0.633	0.253	0.247

Table 2: The time of reenactment is measured in FPS. All results exclude the time for getting driving parameters that can be calculated in advance and are averaged over 100 frames.

	StyleHeat	ROME	OTAvatar	HideNeRF	GOHA	CVTHead	GPAvatar	Real3D	P4D	P4D-v2	Ours
Driving FPS	19.82	11.21	0.12	9.73	6.57	18.09	16.86	4.55	9.49	9.62	67.12

Results

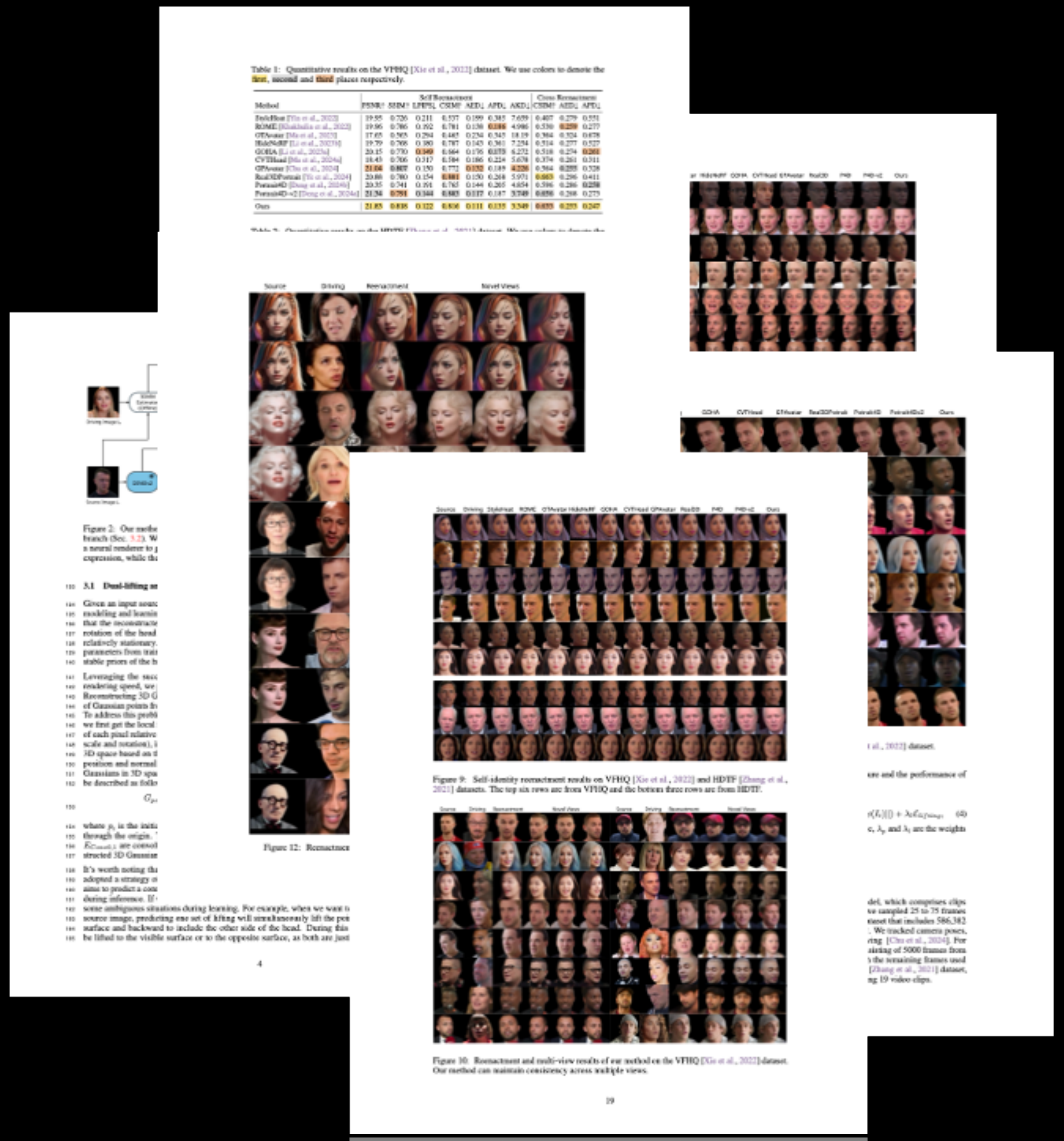
These visualizations demonstrate the generality and robustness of our approach across various inputs, driving poses and expressions.

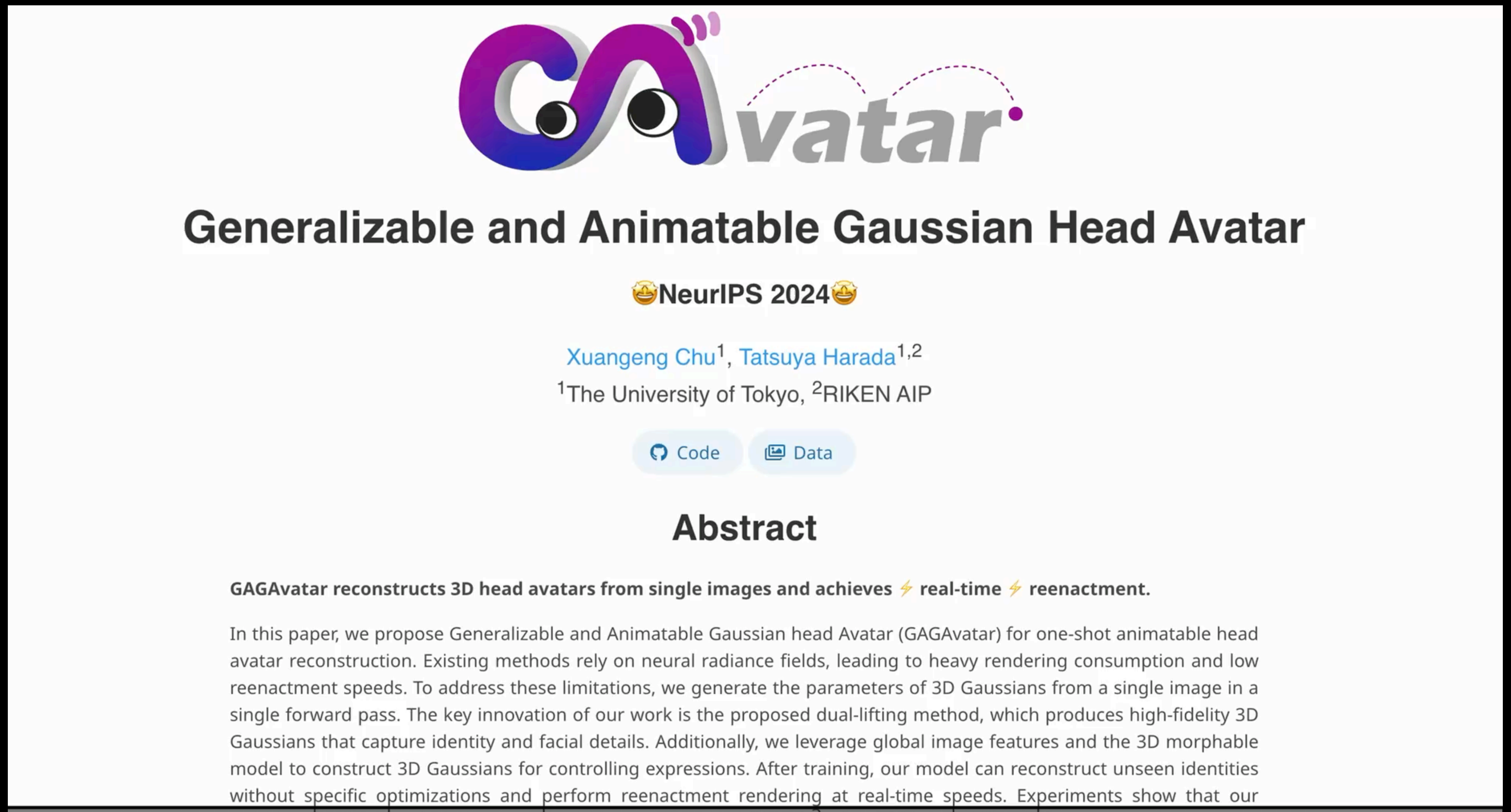
Results

Our approach provides video stability without further processing.

Results

More results can be found in paper and project website.





Generalizable and Animatable Gaussian Head Avatar

Xuangeng Chu[1], Tatsuya Harada[1,2]

The University of Tokyo[1], RIKEN AIP[2]