SPATIAL-QUERY MULTIMODAL ROUTING FOR AUDIO-DRIVEN 3D GAUSSIAN TALKING HEADS

Anonymous authors
Paper under double-blind review

ABSTRACT

Audio-driven 3D talking heads require a representation that maps global speech signals to localized facial motion while preserving identity and rendering fidelity. Although spatial-query conditioning has been adopted in dynamic 3D Gaussian talking heads, its interaction with local reconstruction objectives and its consistency across identities remain underexplored. Building on GaussianTalker, we study a dynamic 3D Gaussian talking-head framework based on Gaussian-wise spatial-query conditioning. Multi-scale planar features encode the spatial context of each primitive, while a Spatial-Audio Cross-Attention (SACA) module uses these features as queries to route information from audio, eye state, camera pose, and a learnable null condition. The resulting Gaussian-specific conditioning predicts residual updates to position, scale, rotation, and appearance, enabling speech-driven deformation to vary across facial regions. A progressive training schedule first optimizes global reconstruction and Gaussian structure and then refines local non-rigid deformation with lip-region reconstruction and foreground-mask supervision. Under an aligned-frame evaluation protocol across three identities, the full configuration achieves the highest macro-averaged PSNR and SSIM among the compared configurations, reaching 33.52 dB and 0.936, respectively. In a 36-run 2-by-2 factorial study of SACA and lip supervision, nine matched Full/No-SACA comparisons show that the full model reduces mean LSE-D by 0.411 and increases mean LSE-C by 0.744, with a mean PSNR decrease of 0.181 dB. Across six directed source-to-target identity pairs, ArcFace cosine similarity ranges from 0.910 to 0.947. These results show that the empirical benefit of the complete spatial-query fusion configuration depends on local supervision and identity, improving average audiovisual synchronization under lip supervision while incurring a measurable reconstruction trade-off.

1 INTRODUCTION

Audio-driven 3D talking heads transform speech into coordinated facial motion and photorealistic renderings (Lahiri et al., 2021; Thies et al., 2020). A shared speech signal must produce spatially distinct responses: the lips articulate phonemes, the eyes follow control signals, head pose changes projection, and other facial regions preserve stable appearance. Spatially selective conditioning is therefore central to the task. 3D Gaussian Splatting (3DGS) represents geometry and appearance through explicit, updatable primitives (Kerbl et al., 2023), making each Gaussian a natural unit for localized multimodal fusion.

Recent systems implement spatial conditioning through region attention and spatial–audio cross-attention. ER-NeRF uses region attention for location-dependent features, and GaussianTalker combines multi-resolution tri-planes with Spatial-Audio Cross-Attention (SACA) for Gaussian residual prediction (Li et al., 2023; Cho et al., 2024). Three factors shape the evaluation of this fusion mechanism. Local lip supervision can change its optimization effect, identity averaging can conceal divergent outcomes, and synchronization gains can coincide with changes in regional reconstruction. Together, these factors shape the observed advantage of the complete fusion configuration.

We investigate SACA as *spatial-query multimodal routing*: each Gaussian-specific spatial feature queries audio, eye state, camera pose, and a learnable null token. A fixed-weight mean-fusion reference uses the same four condition tokens and spatial feature. Crossing fusion with lip supervision

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

yields a 2×2 factorial study over three identities and three seeds (36 runs), quantifying how local supervision changes the relative complete-module effect. Fifteen additional identities extend the lip-on comparison at seed 1 and characterize variation across identity–video instances.

The results establish three main findings. **(1) Local supervision changes the average synchronization effect:** with lip supervision, SACA improves mean LSE-D by 0.411 and LSE-C by 0.744 relative to mean fusion; without lip supervision, both directions reverse. **(2) Effects vary across identities:** among fifteen additional seed-1 identities, nine improve on both synchronization metrics, two show mixed changes, and four worsen on both. **(3) Synchronization improvements coincide with a regional reconstruction cost:** mouth PSNR decreases for all fifteen identities in the seed-1 extension, with a mean change of -0.307 dB. The study contributes a factorial analysis of complete-module fusion, an identity-resolved extension, and a joint evaluation of audiovisual synchronization and regional reconstruction.

2 Related work

2.1 Speech representations and 3D talking heads

Self-supervised models such as wav2vec 2.0 and HuBERT learn contextual speech features through masked contrastive learning and prediction of offline cluster assignments, respectively (Baevski et al., 2020; Hsu et al., 2021). Talking-head systems connect speech to appearance through different motion representations: Neural Voice Puppetry predicts latent 3D motion, LipSync3D learns geometry and texture with pose/lighting normalization, and SadTalker predicts expression and pose from a reference image (Thies et al., 2020; Lahiri et al., 2021; Zhang et al., 2023).

Within neural rendering, NeRF's coordinate-based representation (Mildenhall et al., 2020) supports audio conditioning in AD-NeRF, audio–spatial decomposition in RAD-NeRF, and tri-plane hashing with region attention in ER-NeRF (Guo et al., 2021; Tang et al., 2022; Li et al., 2023). Instant-NGP improves representation efficiency through hash encoding (Müller et al., 2022). GeneFace combines variational motion generation, target adaptation, and NeRF rendering; GeneFace++ incorporates pitch and temporal constraints (Ye et al., 2023b;a). SyncTalk addresses consistency across lips, expressions, pose, and compositing (Peng et al., 2024). Our evaluation considers both audiovisual synchronization and reconstruction because they measure complementary aspects of these systems.

2.2 Dynamic Gaussians and spatial conditioning

Planar representations provide spatial features for neural rendering: EG3D uses tri-planes, K-Planes factorizes space, time, and appearance, and HexPlane represents dynamic scenes with six planes (Chan et al., 2022; Fridovich-Keil et al., 2023; Cao & Johnson, 2023). GaussianTalker combines multi-scale spatial planes with 3DGS primitives and uses spatial queries to fuse audio and auxiliary conditions for residual prediction (Kerbl et al., 2023; Cho et al., 2024).

Dynamic Gaussian methods address motion and adaptation through several designs. TalkingGaussian uses structure-persistent deformation with separate face and mouth-interior representations, while 4D-GS predicts deformation from spatiotemporal features (Li et al., 2024; Wu et al., 2024). InsTaG and GGTalker use motion or cross-modal priors for identity adaptation (Li et al., 2025; Hu et al., 2025); GaussianHeadTalk uses a morphable-model Gaussian mapping for temporal stability (Agarwal et al., 2026).

Conditioning also spans feature-wise affine modulation in FiLM and temporal cross-modal attention in MulT (Perez et al., 2018; Tsai et al., 2019). In the configuration studied here, spatial queries attend to four frame-level conditions. The mean-fusion replacement provides a fixed-weight reference for this Gaussian-wise fusion module.

2.3 Local objectives and configuration evaluation

Local objectives connect generated motion to different supervision signals. Wav2Lip uses a pre-trained synchronization expert, while GaussianTalker adds paired lip-image reconstruction during deformation training (Prajwal et al., 2020; Cho et al., 2024). We study the latter objective alongside the global image and perceptual losses (Johnson et al., 2016). Crossing fusion with lip supervision

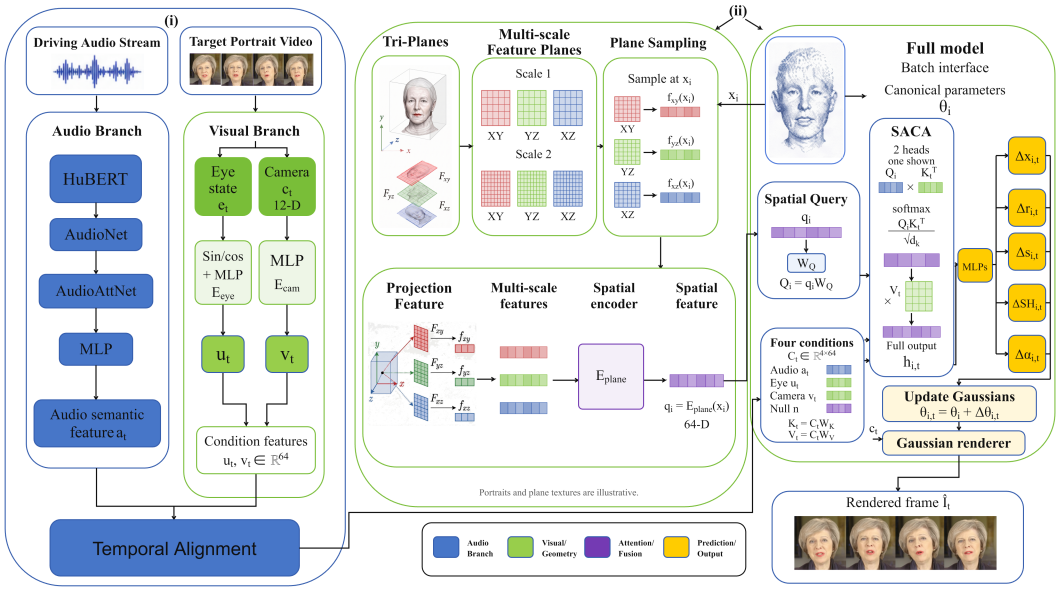


Figure 1: Evaluated GaussianTalker-based SACA pipeline. Gaussian-specific spatial features query four frame-level conditions, and the fused representation predicts five Gaussian attribute residuals; one of two attention heads is illustrated. Appendix C.3 documents configuration correspondence.

permits a difference-in-differences analysis of their joint effect. Following evidence that initialization and data variation affect benchmark comparisons (Bouthillier et al., 2021), we evaluate paired seeds and report identity-specific outcomes.

3 Studied Framework and Factorial Design

3.1 Task and Overview

For each target identity, we learn a canonical Gaussian representation G_0 and an audio-conditioned deformation model from paired video and audio. A sample contains a reference frame I_t, aligned audio context A_t, eye control e_t, and camera input c_t. Conditional attribute residuals produce a time-dependent Gaussian set G_t, rendered as $\hat{I}_t$.

Figure 1 shows the evaluated GaussianTalker-based pipeline (Cho et al., 2024). We retain its spatial and condition encoders, SACA module, Gaussian updates, and staged training, and evaluate the interaction between complete-module fusion and lip supervision. Appendix B.4 records component provenance.

3.2 Spatial Features and Multimodal Conditions

At canonical Gaussian center x_i, shared learnable XY/YZ/XZ planes yield a 64-dimensional spatial feature. The base resolution is 64, scale factors are $\{1, 2\}$, and plane feature channels are 32. AudioNet encodes 1024-dimensional HuBERT features (Hsu et al., 2021); AudioAttNet aggregates their temporal context, followed by a mapping to 64 dimensions. Eye control uses a 32-dimensional sine/cosine encoding and an MLP; a separate MLP maps the 12-dimensional camera input to 64 dimensions:

$$q_i = E_{\text{plane}}(x_i), \qquad a_t = E_{\text{audio}}(A_t), \tag{1}$$

$$u_t = E_{\text{eye}}(e_t), \qquad v_t = E_{\text{cam}}(c_t). \tag{2}$$

A learned 64-dimensional null vector n is shared across frames. Stacking these conditions gives

$$C_t = \text{Stack}(a_t, u_t, v_t, n) \in \mathbb{R}^{4\times64}. \tag{3}$$

All Gaussians in frame t share C_t and have position-dependent q_i. Temporal audio context is aggregated into a single token before SACA.

3.3 Spatial-query fusion and Gaussian updates

For one scaled dot-product attention head (Vaswani et al., 2017), the core computation is

$$Q_i = q_i W_Q, \qquad K_t = C_t W_K, \qquad V_t = C_t W_V, \tag{4}$$

$$w_{i,t} = \mathrm{softmax}\Big(Q_i K_t^\top / \sqrt{d_k}\Big), \qquad z_{i,t} = w_{i,t} V_t. \tag{5}$$

Softmax normalizes over the four conditions, producing $w_{i,t} \in \mathbb{R}^{1\times 4}$. The complete module, including subsequent transformations, produces

$$h_{i,t} = T_{\mathrm{SACA}}(q_i, C_t) \in \mathbb{R}^{64}. \tag{6}$$

Here, $z_{i,t}$ denotes one head's weighted value vector and $h_{i,t}$ the full module output. The evaluated configuration uses one layer, two heads, a feed-forward width of 128, and dropout 0.2.

Five independent two-linear-layer heads, with ReLU and hidden width 128, predict position (3), scale (3), quaternion rotation (4), opacity (1), and spherical-harmonic appearance (16×3) residuals. Let θ_i denote the base parameters at the deformation interface and $\Delta\theta_{i,t}$ the combined residuals. Then

$$\theta_{i,t} = \theta_i + \Delta\theta_{i,t}, \qquad \hat{I}_t = R(G_t, c_t). \tag{7}$$

The addition operates in the module's parameter representation; R includes conversion to rendering attributes and Gaussian rasterization.

3.4 Two-stage training and objectives

The global objective combines image L_1, VGG feature reconstruction (Johnson et al., 2016; Simonyan & Zisserman, 2015), and structural similarity (Wang et al., 2004):

$$\mathcal{L}_{\mathrm{global}} = 0.8\mathcal{L}_{\mathrm{image}} + 0.01\mathcal{L}_{\mathrm{VGG}} + 0.2[1 - \mathrm{SSIM}(\hat{I}_t, I_t)]. \tag{8}$$

During refinement, paired lip crops supply $\mathcal{L}_{\mathrm{lip}}$. For the foreground term, D_t denotes the renderer-interface output used by the training implementation and M_t the target mask:

$$\mathcal{L}_{\mathrm{lip}} = L_1(\hat{I}_t^{\mathrm{lip}}, I_t^{\mathrm{lip}}), \tag{9}$$

$$\hat{M}_t = 2\,\mathrm{sigmoid}(D_t) - 1, \qquad \mathcal{L}_{\mathrm{mask}} = L_1(\hat{M}_t, M_t), \tag{10}$$

$$\mathcal{L}_{\mathrm{fine}} = \mathcal{L}_{\mathrm{global}} + 0.4\mathcal{L}_{\mathrm{lip}} + 0.4\mathcal{L}_{\mathrm{mask}}. \tag{11}$$

All configurations train for 8,000 global iterations at batch size 1 and 2,000 refinement iterations at batch size 32. Lip-off configurations omit only $0.4\mathcal{L}_{\mathrm{lip}}$, retaining foreground-mask supervision. Densification and pruning follow the coarse-stage schedule; densification and splitting are disabled during refinement.

3.5 Factorial design: fusion and lip supervision

Each run corresponds to an identity, seed, and configuration. Both branches retain the spatial representation, four input types, residual-head architecture, and two-stage schedule, but optimize their parameters separately. We refer to the two branches as *SACA fusion* and *mean fusion* in the text; tables retain the compact Full and No-SACA configuration labels. The mean branch averages condition tokens, adds the spatial input, and applies LayerNorm (Ba et al., 2016):

$$\bar{c}_t = \tfrac{1}{4}(a_t + u_t + v_t + n), \qquad h_{i,t}^{\mathrm{mean}} = \mathrm{LayerNorm}(q_i + \bar{c}_t). \tag{12}$$

The mean branch remains spatially dependent through q_i. This *complete-module replacement* jointly changes condition weighting, active capacity, feed-forward computation, and gradient paths.

For metric Y, within each identity–seed pair we compute

$$\delta_{\mathrm{on}} = Y(\text{Full}) - Y(\text{No-SACA}), \tag{13}$$

$$\delta_{\mathrm{off}} = Y(\text{No-Lip}) - Y(\text{No-SACA-No-Lip}), \tag{14}$$

$$\delta_{\mathrm{interaction}} = \delta_{\mathrm{on}} - \delta_{\mathrm{off}}. \tag{15}$$

Table 1: Factorial configurations. Local terms apply during refinement. The mean branch replaces the complete SACA module (Section 3.5).

Configuration	Fusion	Lip term	Mask term
Full	SACA	$0.4\mathcal{L}_{\text{lip}}$	$0.4\mathcal{L}_{\text{mask}}$
No-SACA	Mean	$0.4\mathcal{L}_{\text{lip}}$	$0.4\mathcal{L}_{\text{mask}}$
No-Lip	SACA	Off	$0.4\mathcal{L}_{\text{mask}}$
No-SACA-No-Lip	Mean	Off	$0.4\mathcal{L}_{\text{mask}}$

We average seeds within identity, then weight identities equally. Negative LSE-D changes and positive LSE-C/PSNR changes are favorable. A favorable interaction means that lip supervision improves the *relative difference*; superiority in a particular setting depends on its conditional effect, δ_{on} or δ_{off}.

4 Experiments

4.1 Evaluation setup

Data and comparisons. The core study covers May, Obama, and Kjell, with three seeds and four configurations (36 runs). Its whole-image validation sets contain 553, 728, and 238 frames, respectively. Fifteen additional identities comprise three HDTF-derived identities (Zhang et al., 2021), two exploratory public-video identities, and ten further public-video identities. Approximately 120-second clips are cropped/resized to 512×512 at 25 fps, with HuBERT audio features. Each added identity has 273 whole-image evaluation frames and Full/No-SACA runs at seed 1 (30 additional runs). All evaluated identities are retained. The 18-identity summary uniformly uses seed 1; core three-seed results remain separate.

External comparisons evaluate available GaussianTalker, GeneFace++, and ER-NeRF checkpoints (Cho et al., 2024; Ye et al., 2023a; Li et al., 2023) on common frame prefixes: May 548, Obama 724, and Kjell 238, at seed 1. These are system-level comparisons with aligned evaluation frames and method-specific training budgets and processing. Appendix B.2 records the settings. The external GaussianTalker checkpoint and the internal mean-fusion branch are separate configurations.

Metrics and aggregation. We report whole-image and mouth PSNR, SSIM (Wang et al., 2004), and AlexNet LPIPS (Zhang et al., 2018); LPIPS is an evaluation metric, while the training objective uses VGG features. Mouth evaluation uses reference lip landmarks, a $1.5\times$ bounding box, and the same crop for prediction and reference. Excluding failed detections leaves 236 valid mouth frames for Kjell. SyncNet-based LSE-D and LSE-C evaluate generated clips (Chung & Zisserman, 2016; Prajwal et al., 2020); ArcFace cosine similarity measures target-identity appearance under cross-source audio (Deng et al., 2019). Reconstruction metrics average valid frames within each trained model, then seeds within identity, and finally identities equally.

Uncertainty. The core study has three identity clusters, each with three seeds. Frames are nested within trained models; statistical replication occurs at the identity and training-seed levels. We use 20,000 hierarchical bootstrap replicates (Efron, 1979), resampling identities and then seeds within identity while preserving four-configuration pairing. Percentile 95% intervals provide exploratory uncertainty summaries for these three identities. Leave-one-identity-out recomputation assesses sensitivity to their composition. Added identities provide single-seed means and direction counts for the lip-on comparison.

4.2 Lip supervision changes the synchronization effect

Lip supervision reverses the average synchronization difference between SACA and mean fusion. With lip supervision, the nine core Full/No-SACA pairs yield mean changes of -0.411 LSE-D and $+0.744$ LSE-C (Table 2). Without it, SACA versus mean fusion yields $+0.301$ LSE-D and -0.076 LSE-C. The interaction estimates are -0.712 and $+0.820$, respectively (Table 3; Figure 2a).

Table 2: Core factorial macro-averages: three identities and three seeds per configuration. PSNR is in dB.

Configuration	PSNR ↑	SSIM ↑	LPIPS ↓	LSE-D ↓	LSE-C ↑
Full	33.529	0.9359	0.0479	11.224	2.816
No-SACA	33.710	0.9378	0.0462	11.635	2.072
No-Lip	33.244	0.9337	0.0463	12.269	0.824
No-SACA-No-Lip	33.559	0.9367	0.0451	11.968	0.900

Table 3: Conditional configuration effects and interaction in original metric directions. Intervals are exploratory hierarchical bootstrap percentiles over three identities.

Metric	δ_{on}	δ_{off}	$\delta_{\text{interaction}}$	Interaction 95% interval
LSE-D	-0.411	+0.301	-0.712	[-1.418, -0.154]
LSE-C	+0.744	-0.076	+0.820	[+0.120, +1.733]
PSNR	-0.181	-0.315	+0.134	[+0.070, +0.183]

The positive PSNR interaction reflects a *smaller reconstruction cost*: SACA–mean PSNR is -0.181 dB with lip supervision and -0.315 dB without it. Lip supervision reduces this gap by 0.134 dB, while mean fusion retains higher PSNR in both settings.

4.3 Identity-dependent synchronization outcomes

Synchronization effects vary across identities. The three-seed paired LSE-D/LSE-C changes are -1.247 ± 1.068/$+1.616 \pm 1.110$ for May, -0.145 ± 0.174/$+0.449 \pm 0.490$ for Obama, and $+0.158 \pm 0.215$/$+0.167 \pm 0.058$ for Kjell (sample SD across paired seed differences). May and Obama improve on both means; Kjell has mixed lip-on outcomes despite favorable interaction point estimates. May also shows the largest synchronization gains and seed variation.

Leaving out May, Obama, or Kjell yields lip-on LSE-D effects of $+0.006$, -0.544, and -0.696, and LSE-C effects of $+0.308$, $+0.892$, and $+1.032$. Distance improvement disappears without May, whereas interaction directions remain favorable in all leave-out combinations (Appendix A.5).

Across the fifteen additional identities, mean LSE-D changes by -0.136 and LSE-C by $+0.264$: nine improve on both, two are mixed, and four worsen on both. Across all eighteen seed-1 pairs, the changes are -0.242 and $+0.384$, with counts 10/3/5 (Table 4; Figure 2b). Group averages also vary: the HDTF-derived group worsens on average, whereas seven of the ten further public-video identities improve on both metrics.

4.4 Synchronization–reconstruction trade-offs

Reconstruction changes vary by region. For May/Obama/Kjell, the core lip-on whole-image PSNR changes are $-0.222/-0.213/-0.107$ dB, while mouth PSNR changes are $+0.020/-0.122/-0.398$ dB. Macro-SSIM decreases by approximately 0.0018 and LPIPS increases by approximately 0.0017 (Appendix A.2).

The extended evaluation reveals a consistent local reconstruction cost: mouth PSNR decreases for *all fifteen* additional identities, averaging -0.307 dB, including all nine with improved synchronization on both metrics. Across eighteen identities, mean whole-image and mouth PSNR changes are -0.213 and -0.276 dB. SSIM and LPIPS changes are reported in Appendix A.3. Under the evaluated loss weights, better clip-level synchronization frequently accompanies lower similarity to the reference mouth images.

4.5 External aligned-frame comparison

In the aligned-frame comparison (Table 5), Full obtains the highest macro-PSNR (33.516 dB) and SSIM (0.9355), exceeding the evaluated GaussianTalker checkpoint by 0.591 dB and 0.0107. Full has the highest PSNR for May and Kjell and the highest SSIM for all three identities (Appendix A.1).

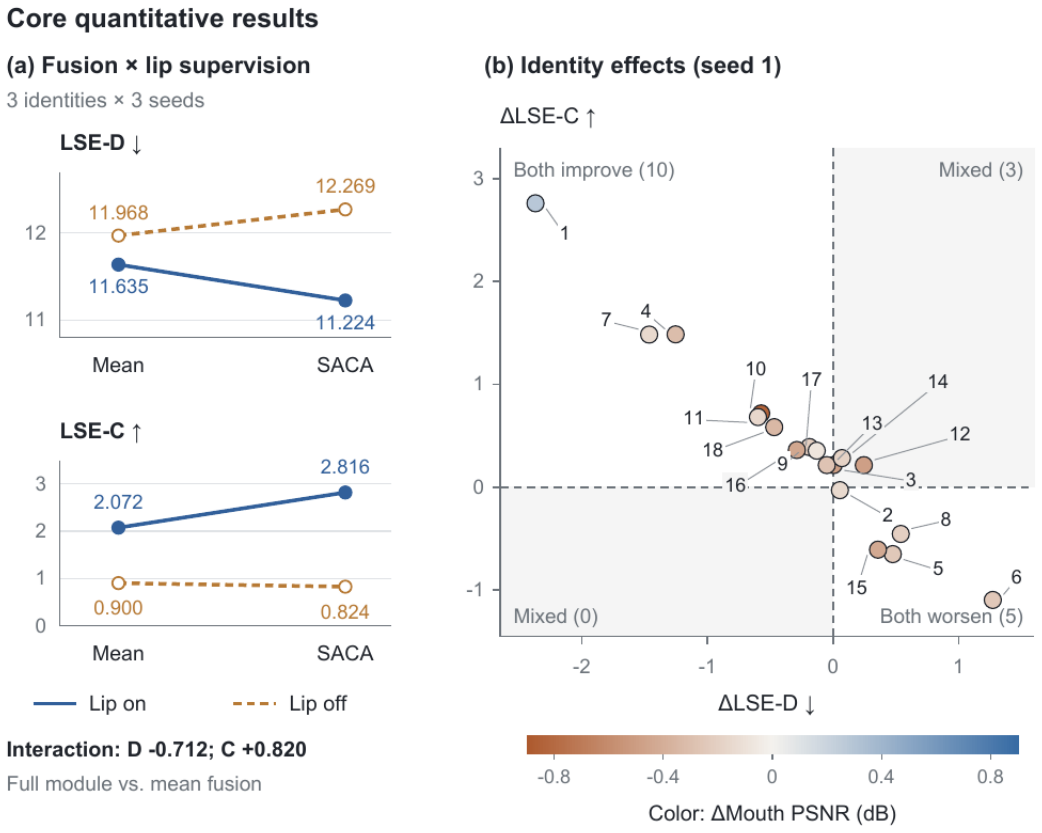


Figure 2: (a) Three-identity, three-seed configuration means; intervals appear in Table 3. (b) Eighteen lip-on, seed-1 Full–No-SACA pairs: 10 improve on both synchronization metrics, 3 are mixed, and 5 worsen. Zero lines mark changes in metric direction. Color encodes mouth PSNR change on a symmetric $[-0.9, +0.9]$ dB scale; points use the original, unjittered coordinates.

Table 4: Seed-1 Full–No-SACA differences with lip supervision. D/C denote LSE-D/LSE-C; M-PSNR is mouth PSNR (dB). B/M/W counts both improved, mixed, and both worsened synchronization. Added 15 combines HDTF 3, Public 2, and Public 10; All 18 also includes Core 3.

Group	n	ΔD	ΔC	ΔPSNR	ΔM-PSNR	B/M/W
Core 3	3	-0.770	+0.983	-0.120	-0.118	1/1/1
HDTF 3	3	+0.165	-0.087	+0.075	-0.269	1/0/2
Public 2	2	-0.462	+0.515	-0.503	-0.178	1/0/1
Public 10	10	-0.162	+0.320	-0.270	-0.345	7/2/1
Added 15	15	-0.136	+0.264	-0.232	-0.307	9/2/4
All 18	18	-0.242	+0.384	-0.213	-0.276	10/3/5

Rankings differ by metric: GaussianTalker has the lowest macro-LPIPS, and GeneFace++ has the best synchronization scores.

Table 5: Aligned-frame external comparison at seed 1. Training and processing are method-specific (Appendix B.2). PSNR is in dB.

Method	PSNR ↑	SSIM ↑	LPIPS ↓	LSE-D ↓	LSE-C ↑
GaussianTalker	32.925	0.9248	0.0410	10.505	4.417
GeneFace++	22.903	0.7920	0.1806	7.313	5.868
ER-NeRF	24.875	0.8345	0.1185	11.420	2.805
Full	33.516	0.9355	0.0475	11.197	2.896

4.6 Identity similarity under cross-source audio

Across the six directed source–target pairs among May, Obama, and Kjell, mean ArcFace cosine similarity to target references ranges from 0.910 to 0.947. Each direction changes the driving audio while retaining the target's separately trained Full/seed-1 model. This evaluation describes known-target identity appearance under changed audio; paired cross-source baselines are unavailable. Appendix A.4 provides per-direction results, sampling, and detection exclusions.

4.7 Qualitative outcomes across effect directions

Figure 3 reports seed-1 cases with improved (May), mixed (Melania Trump), and worsened (Lauren Underwood) synchronization, using matched output positions and mouth crops. May's synchronization and mouth PSNR improvements coexist with a 0.127 dB whole-image PSNR decrease. For Lauren, the reference mouth is nearly closed while Full remains visibly open; No-SACA more closely matches the reference opening.

Other May positions still deviate from the reference, illustrating variation within a sequence that improves on average.

5 Discussion and limitations

The dependence on lip supervision makes the training objective part of the fusion comparison. Adding local reconstruction changes the relative synchronization performance of the two fusion configurations, motivating evaluation across local-loss settings. To investigate how this complete-module effect arises, future experiments could perturb spatial queries and condition tokens while matching active capacity and preserving feed-forward paths. Work on attention interpretability in NLP similarly motivates pairing weight inspection with interventions (Jain & Wallace, 2019).

The separation between synchronization and mouth reconstruction matters for model selection. Selecting a configuration by mean synchronization alone can favor outputs with larger mouth-image errors. Identity-resolved synchronization and regional reconstruction scores make this choice more explicit. The evaluated outputs also contain torso/lectern defects for Joe Biden and misalignment, multiple reference faces, and local defects in Mike Pence examples; both identities remain in the statistics. Audiovisual inspection and subjective evaluation would help relate these visual errors and metric changes to viewing quality.

The paired-seed core design estimates the fusion–supervision interaction across three identities, and the fifteen-identity extension broadens the observed lip-on direction patterns at seed 1. Broader factorial replication and local-loss weight sweeps remain valuable directions for assessing the stability of the interaction and the synchronization–reconstruction trade-off. The supplementary archive supports verification of the reported tables, plots, and analyses; Appendix B.3 defines the release scope.

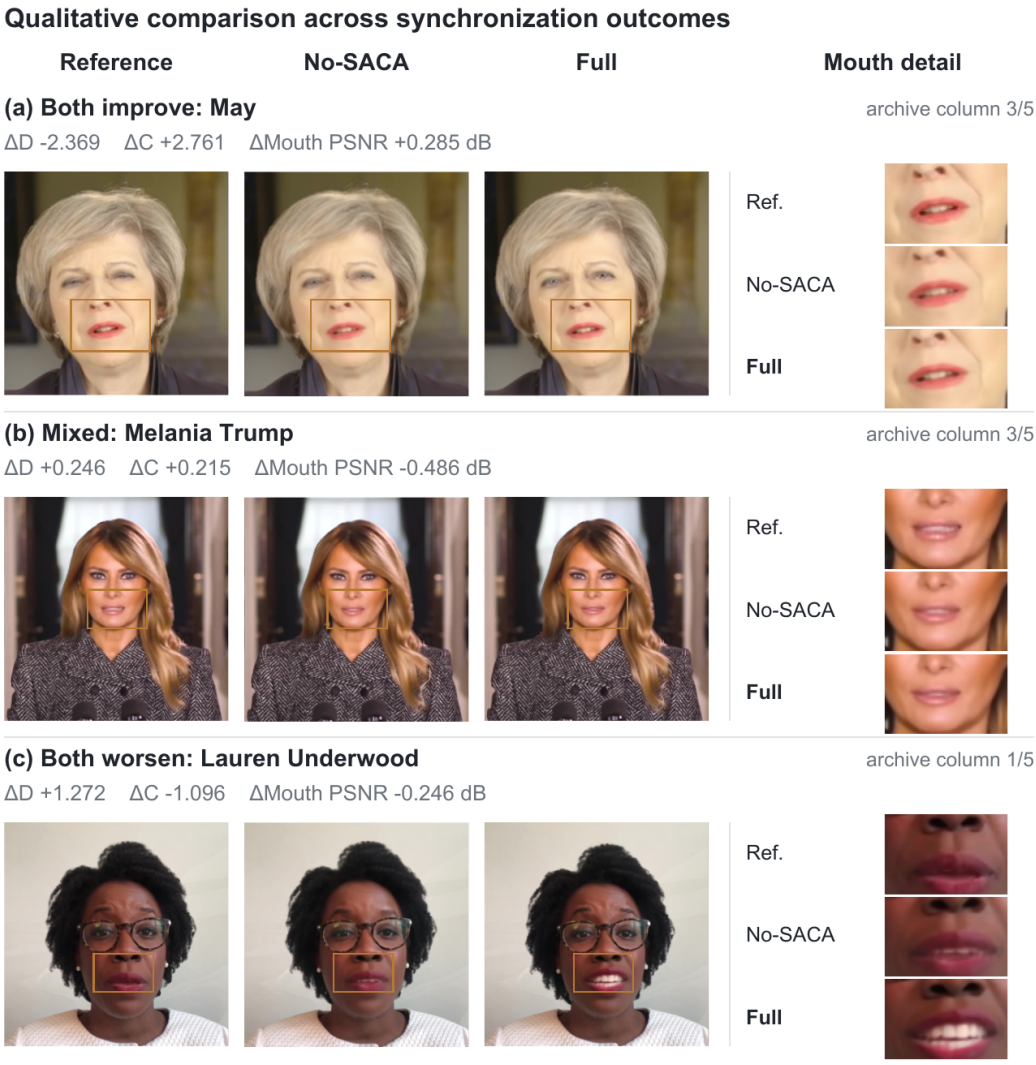


Figure 3: Examples selected for improved, mixed, and worsened sequence-level synchronization. Mouth enlargements retain the reference/No-SACA/Full order and identical windows. Reported differences are sequence-level Full–No-SACA scores. Archive positions are marked; Appendix C.2 documents selection and crops.

6 Conclusion

Our factorial study of GaussianTalker-based talking heads identifies a supervision-dependent synchronization effect of the complete SACA module. Lip reconstruction changes its average ranking against mean fusion, while identity-resolved evaluation reveals improved, mixed, and worsened synchronization. In the extended lip-on comparison, average synchronization gains accompany lower mouth PSNR. Together, these findings support assessing fusion configurations jointly with local supervision and reporting synchronization alongside regional reconstruction across identities.

ETHICS STATEMENT

Audio-driven avatars can support virtual communication, accessibility, and content creation, while also enabling impersonation and misleading audiovisual content. Use and redistribution of source videos, audio, identity images, and derivative assets must follow applicable permissions and licenses. Generated media should be clearly labeled as synthetic, and deployment should include provenance and authorization checks. Appendix C documents the sources of illustrative assets and experimental images.

REPRODUCIBILITY STATEMENT

Sections 3.2–3.5 define the interfaces, objectives, and paired comparisons. Appendix A gives identity-level and run-level results and sensitivity analyses; Appendix B records architecture and comparison settings; Appendix C documents plotting and image selection. The supplementary archive supports regeneration of the reported tables, plots, and analyses from recorded outputs. End-to-end retraining additionally depends on the complete GaussianTalker-based implementation and its preprocessing environment; Appendix B.3 defines the release scope.

AI USE STATEMENT

Generative AI tools were used to aid and polish writing; support literature retrieval and discovery, including finding related work; assist with research ideation and execution; and draft portions of the manuscript. Specifically, the tools assisted with language revision and manuscript organization, identifying potentially relevant literature, discussing and refining research questions and experimental procedures, interpreting existing experimental results, and preparing initial prose that was subsequently reviewed and revised by the authors. Generative AI tools were not used to generate synthetic datasets. All AI-assisted material was independently checked by the authors against primary sources and experimental records. The authors take full responsibility for the paper's text, citations, scientific claims, and final submission.

REFERENCES

Madhav Agarwal, Mingtian Zhang, Laura Sevilla-Lara, and Steven McDonagh. GaussianHeadTalk: Wobble-Free 3D Talking Heads with Audio Driven Gaussian Splatting. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 8017–8027, 2026. URL https://madhav1ag.github.io/gaussianheadtalk/.

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. Layer Normalization. *arXiv preprint arXiv:1607.06450*, 2016. URL https://arxiv.org/abs/1607.06450.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. In *Advances in Neural Information Processing Systems*, volume 33, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/92d1e1eb1cd6f9fba3227870bb6d7f07-Abstract.html.

Xavier Bouthillier, Pierre Delaunay, Mirko Bronzi, Assya Trofimov, Brennan Nichyporuk, Justin Szeto, Nazanin Mohammadi Sepahvand, Edward Raff, Kanika Madan, Vikram Voleti, Samira Ebrahimi Kahou, Vincent Michalski, Tal Arbel, Chris Pal, Gael Varoquaux, and Pascal Vincent. Accounting for Variance in Machine Learning Benchmarks. *Proceedings of Machine Learning and Systems*, 3, 2021. URL https://proceedings.mlsys.org/paper_files/paper/2021/hash/0184b0cd3cfb185989f858a1d9f5c1eb-Abstract.html.

Ang Cao and Justin Johnson. HexPlane: A Fast Representation for Dynamic Scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 130–141, 2023. URL https://openaccess.thecvf.com/content/CVPR2023/html/Cao_HexPlane_A_Fast_Representation_for_Dynamic_Scenes_CVPR_2023_paper.html.

Eric R. Chan, Connor Z. Lin, Matthew A. Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas J. Guibas, Jonathan Tremblay, Sameh Khamis, Tero Karras, and Gordon Wetzstein. Efficient Geometry-Aware 3D Generative Adversarial Networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16123–16133, 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/Chan_Efficient_Geometry-Aware_3D_Generative_Adversarial_Networks_CVPR_2022_paper.html.

Kyusun Cho, Joungbin Lee, Heeji Yoon, Yeobin Hong, Jaehoon Ko, Sangjun Ahn, and Seungryong Kim. GaussianTalker: Real-Time High-Fidelity Talking Head Synthesis with Audio-Driven 3D Gaussian Splatting. *arXiv preprint arXiv:2404.16012*, 2024. URL https://arxiv.org/abs/2404.16012.

Joon Son Chung and Andrew Zisserman. Out of time: automated lip sync in the wild. In *Workshop on Multi-view Lip-reading, Asian Conference on Computer Vision*, 2016. URL https://www.robots.ox.ac.uk/~vgg/publications/2016/Chung16a/.

Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. ArcFace: Additive Angular Margin Loss for Deep Face Recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4690–4699, 2019. URL https://openaccess.thecvf.com/content_CVPR_2019/html/Deng_ArcFace_Additive_Angular_Margin_Loss_for_Deep_Face_Recognition_CVPR_2019_paper.html.

Bradley Efron. Bootstrap Methods: Another Look at the Jackknife. *The Annals of Statistics*, 7 (1):1–26, 1979. doi: 10.1214/aos/1176344552. URL https://doi.org/10.1214/aos/1176344552.

Sara Fridovich-Keil, Giacomo Meanti, Frederik Rahbæk Warburg, Benjamin Recht, and Angjoo Kanazawa. K-Planes: Explicit Radiance Fields in Space, Time, and Appearance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12479–12488, 2023. URL https://openaccess.thecvf.com/content/CVPR2023/html/Fridovich-Keil_K-Planes_Explicit_Radiance_Fields_in_Space_Time_and_Appearance_CVPR_2023_paper.html.

Yudong Guo, Keyu Chen, Sen Liang, Yong-Jin Liu, Hujun Bao, and Juyong Zhang. AD-NeRF: Audio Driven Neural Radiance Fields for Talking Head Synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5784–5794, 2021. URL https://openaccess.thecvf.com/content/ICCV2021/html/Guo_AD-NeRF_Audio_Driven_Neural_Radiance_Fields_for_Talking_Head_Synthesis_ICCV_2021_paper.html.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units. *arXiv preprint arXiv:2106.07447*, 2021. URL https://arxiv.org/abs/2106.07447.

Wentao Hu, Shunkai Li, Ziqiao Peng, Haoxian Zhang, Fan Shi, Xiaoqiang Liu, Pengfei Wan, Di Zhang, and Hui Tian. GGTalker: Talking Head Systhesis with Generalizable Gaussian Priors and Identity-Specific Adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10108–10117, 2025. URL https://openaccess.thecvf.com/content/ICCV2025/html/Hu_GGTalker_Talking_Head_Systhesis_with_Generalizable_Gaussian_Priors_and_Identity-Specific_ICCV_2025_paper.html.

Sarthak Jain and Byron C. Wallace. Attention is not Explanation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, volume 1, pp. 3543–3556, 2019. doi: 10.18653/v1/N19-1357. URL https://aclanthology.org/N19-1357/.

Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual Losses for Real-Time Style Transfer and Super-Resolution. In *European Conference on Computer Vision*, 2016. URL https://cs.stanford.edu/people/jcjohns/eccv16/.

Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3D Gaussian Splatting for Real-Time Radiance Field Rendering. *ACM Transactions on Graphics*, 42(4), 2023. URL https://repo-sam.inria.fr/fungraph/3d-gaussian-splatting/.

Avisek Lahiri, Vivek Kwatra, Christian Frueh, John Lewis, and Chris Bregler. LipSync3D: Data-Efficient Learning of Personalized 3D Talking Faces From Video Using Pose and Lighting Normalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2755–2764, 2021. URL https://openaccess.thecvf.com/content/CVPR2021/papers/Lahiri_LipSync3D_Data-Efficient_Learning_of_Personalized_3D_Talking_Faces_From_Video_CVPR_2021_paper.pdf.

Jiahe Li, Jiawei Zhang, Xiao Bai, Jun Zhou, and Lin Gu. Efficient Region-Aware Neural Radiance Fields for High-Fidelity Talking Portrait Synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7568–7578, 2023. URL https://openaccess.thecvf.com/content/ICCV2023/html/Li_Efficient_Region-Aware_Neural_Radiance_Fields_for_High-Fidelity_Talking_Portrait_Synthesis_ICCV_2023_paper.html.

Jiahe Li, Jiawei Zhang, Xiao Bai, Jin Zheng, Xin Ning, Jun Zhou, and Lin Gu. TalkingGaussian: Structure-Persistent 3D Talking Head Synthesis via Gaussian Splatting. In *European Conference on Computer Vision*, 2024. URL https://www.ecva.net/papers/eccv_2024/papers_ECCV/html/1539_ECCV_2024_paper.php.

Jiahe Li, Jiawei Zhang, Xiao Bai, Jin Zheng, Jun Zhou, and Lin Gu. InsTaG: Learning Personalized 3D Talking Head from Few-Second Video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10690–10700, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Li_InsTaG_Learning_Personalized_3D_Talking_Head_from_Few-Second_Video_CVPR_2025_paper.html.

Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis. In *European Conference on Computer Vision*, 2020. URL https://arxiv.org/abs/2003.08934.

Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant Neural Graphics Primitives with a Multiresolution Hash Encoding. *ACM Transactions on Graphics*, 41(4): 102:1–102:15, 2022. doi: 10.1145/3528223.3530127. URL https://nvlabs.github.io/instant-ngp/.

Ziqiao Peng, Wentao Hu, Yue Shi, Xiangyu Zhu, Xiaomei Zhang, Hao Zhao, Jun He, Hongyan Liu, and Zhaoxin Fan. SyncTalk: The Devil is in the Synchronization for Talking Head Synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 666–676, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Peng_SyncTalk_The_Devil_is_in_the_Synchronization_for_Talking_Head_CVPR_2024_paper.html.

Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron Courville. FiLM: Visual Reasoning with a General Conditioning Layer. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1), 2018. doi: 10.1609/aaai.v32i1.11671. URL https://ojs.aaai.org/index.php/AAAI/article/view/11671.

K. R. Prajwal, Rudrabha Mukhopadhyay, Vinay P. Namboodiri, and C. V. Jawahar. A Lip Sync Expert Is All You Need for Speech to Lip Generation In The Wild. In *Proceedings of the 28th ACM International Conference on Multimedia*, 2020. doi: 10.1145/3394171.3413532. URL https://doi.org/10.1145/3394171.3413532.

Karen Simonyan and Andrew Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. In *International Conference on Learning Representations*, 2015. URL https://www.robots.ox.ac.uk/~vgg/publications/2015/Simonyan15/.

Jiaxiang Tang, Kaisiyuan Wang, Hang Zhou, Xiaokang Chen, Dongliang He, Tianshu Hu, Jingtuo Liu, Gang Zeng, and Jingdong Wang. Real-time Neural Radiance Talking Portrait Synthesis

via Audio-spatial Decomposition. *arXiv preprint arXiv:2211.12368*, 2022. URL https://arxiv.org/abs/2211.12368.

Justus Thies, Mohamed Elgharib, Ayush Tewari, Christian Theobalt, and Matthias Nießner. Neural Voice Puppetry: Audio-driven Facial Reenactment. In *European Conference on Computer Vision*, 2020. URL https://www.ecva.net/papers/eccv_2020/papers_ECCV/html/2619_ECCV_2020_paper.php.

Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J. Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. Multimodal Transformer for Unaligned Multimodal Language Sequences. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 6558–6569, 2019. doi: 10.18653/v1/P19-1656. URL https://aclanthology.org/P19-1656/.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention Is All You Need. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image Quality Assessment: From Error Visibility to Structural Similarity. *IEEE Transactions on Image Processing*, 13(4): 600–612, 2004. URL https://ece.uwaterloo.ca/~z70wang/research/ssim/.

Guanjun Wu, Taoran Yi, Jiemin Fang, Lingxi Xie, Xiaopeng Zhang, Wei Wei, Wenyu Liu, Qi Tian, and Xinggang Wang. 4D Gaussian Splatting for Real-Time Dynamic Scene Rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20310–20320, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Wu_4D_Gaussian_Splatting_for_Real-Time_Dynamic_Scene_Rendering_CVPR_2024_paper.html.

Zhenhui Ye, Jinzheng He, Ziyue Jiang, Rongjie Huang, Jiawei Huang, Jinglin Liu, Yi Ren, Xiang Yin, Zejun Ma, and Zhou Zhao. GeneFace++: Generalized and Stable Real-Time Audio-Driven 3D Talking Face Generation. *arXiv preprint arXiv:2305.00787*, 2023a. URL https://arxiv.org/abs/2305.00787.

Zhenhui Ye, Ziyue Jiang, Yi Ren, Jinglin Liu, JinZheng He, and Zhou Zhao. GeneFace: Generalized and High-Fidelity Audio-Driven 3D Talking Face Synthesis. In *International Conference on Learning Representations*, 2023b. URL https://openreview.net/forum?id=YfwMIDhPccD.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 586–595, 2018. URL https://openaccess.thecvf.com/content_cvpr_2018/html/Zhang_The_Unreasonable_Effectiveness_CVPR_2018_paper.html.

Wenxuan Zhang, Xiaodong Cun, Xuan Wang, Yong Zhang, Xi Shen, Yu Guo, Ying Shan, and Fei Wang. SadTalker: Learning Realistic 3D Motion Coefficients for Stylized Audio-Driven Single Image Talking Face Animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8652–8661, 2023. URL https://openaccess.thecvf.com/content/CVPR2023/html/Zhang_SadTalker_Learning_Realistic_3D_Motion_Coefficients_for_Stylized_Audio-Driven_Single_CVPR_2023_paper.html.

Zhimeng Zhang, Lincheng Li, Yu Ding, and Changjie Fan. Flow-Guided One-Shot Talking Face Generation With a High-Resolution Audio-Visual Dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3661–3670, 2021. URL https://openaccess.thecvf.com/content/CVPR2021/html/Zhang_Flow-Guided_One-Shot_Talking_Face_Generation_With_a_High-Resolution_Audio-Visual_Dataset_CVPR_2021_paper.html.

A SUPPLEMENTARY EXPERIMENTAL RESULTS

This appendix provides identity-level and run-level results for the external comparison, core factorial study, and seed-1 extension.

A.1 EXTERNAL COMPARISONS BY IDENTITY

All methods use the common validation-frame prefix for each identity and archived seed-1 results. Metrics are first aggregated within each evaluated model and then macro-averaged across identities in Table 5. PSNR is measured in dB.

Table 6: External aligned-frame results by identity (seed 1). PSNR is in dB.

Identity	Method	Frames	PSNR ↑	SSIM ↑	LPIPS ↓	LSE-D ↓	LSE-C ↑
May	GaussianTalker	548	28.543	0.8498	0.0791	9.165	5.516
May	GeneFace++	548	26.413	0.7956	0.1490	7.228	7.608
May	ER-NeRF	548	26.529	0.8010	0.0881	11.521	2.810
May	Full	548	28.829	0.8573	0.1008	8.660	4.760
Obama	GaussianTalker	724	35.439	0.9600	0.0252	8.941	6.449
Obama	GeneFace++	724	22.943	0.8764	0.0750	6.210	8.183
Obama	ER-NeRF	724	27.025	0.9299	0.0438	9.931	4.716
Obama	Full	724	35.243	0.9660	0.0268	10.499	2.908
Kjell	GaussianTalker	238	34.794	0.9646	0.0189	13.409	1.285
Kjell	GeneFace++	238	19.353	0.7041	0.3179	8.502	1.813
Kjell	ER-NeRF	238	21.071	0.7727	0.2235	12.809	0.889
Kjell	Full	238	36.475	0.9833	0.0150	14.432	1.020

Method-specific training and processing settings are recorded in Appendix B.2.

A.2 CORE FACTORIAL RESULTS BY IDENTITY

A.2.1 WHOLE-IMAGE RECONSTRUCTION AND SYNCHRONIZATION

Table 7: Core whole-image and synchronization results: three-seed mean and sample SD. Each cell shows the mean above its SD. F: Full; M: No-SACA; L: No-Lip; N: No-SACA-No-Lip. PSNR is in dB.

Identity	Config.	PSNR	SSIM	LPIPS	LSE-D	LSE-C
May	F	28.845 ± 0.037	0.8583 ± 0.0015	0.1017 ± 0.0021	9.234 ± 0.489	4.057 ± 0.587
May	M	29.067 ± 0.106	0.8629 ± 0.0020	0.0973 ± 0.0006	10.481 ± 0.584	2.441 ± 0.557
May	L	28.556 ± 0.110	0.8538 ± 0.0029	0.0971 ± 0.0007	11.675 ± 0.128	1.060 ± 0.108
May	N	28.911 ± 0.090	0.8606 ± 0.0027	0.0948 ± 0.0018	11.645 ± 0.050	1.176 ± 0.135
Obama	F	35.251 ± 0.094	0.9661 ± 0.0002	0.0270 ± 0.0002	10.136 ± 0.244	3.486 ± 0.492
Obama	M	35.464 ± 0.124	0.9665 ± 0.0004	0.0264 ± 0.0005	10.281 ± 0.179	3.037 ± 0.430
Obama	L	34.968 ± 0.076	0.9647 ± 0.0005	0.0271 ± 0.0004	11.664 ± 0.164	0.709 ± 0.069
Obama	N	35.355 ± 0.035	0.9661 ± 0.0003	0.0260 ± 0.0003	11.630 ± 0.103	0.867 ± 0.100
Kjell	F	36.492 ± 0.099	0.9834 ± 0.0003	0.0150 ± 0.0001	14.301 ± 0.221	0.906 ± 0.164
Kjell	M	36.599 ± 0.050	0.9839 ± 0.0002	0.0149 ± 0.0001	14.143 ± 0.144	0.739 ± 0.125

Continued on next page

Table 7 continued

Identity	Config.	PSNR	SSIM	LPIPS	LSE-D	LSE-C
Kjell	L	36.208 ± 0.044	0.9827 ± 0.0002	0.0148 ± 0.0001	13.467 ± 0.149	0.704 ± 0.052
Kjell	N	36.410 ± 0.022	0.9835 ± 0.0000	0.0146 ± 0.0001	12.629 ± 0.315	0.657 ± 0.115

A.2.2 Mouth reconstruction

Valid mouth-frame counts are recorded in Table 15.

Table 8: Core mouth reconstruction: three-seed mean ± sample SD. Configuration codes follow Table 7; PSNR is in dB.

Identity	Config.	Mouth PSNR	Mouth SSIM	Mouth LPIPS
May	F	23.679 ± 0.102	0.7543 ± 0.0008	0.1523 ± 0.0028
May	M	23.659 ± 0.154	0.7549 ± 0.0052	0.1474 ± 0.0039
May	L	22.897 ± 0.083	0.7378 ± 0.0011	0.1448 ± 0.0001
May	N	23.036 ± 0.023	0.7422 ± 0.0014	0.1489 ± 0.0038
Obama	F	26.227 ± 0.167	0.7634 ± 0.0031	0.1565 ± 0.0033
Obama	M	26.349 ± 0.134	0.7665 ± 0.0031	0.1528 ± 0.0033
Obama	L	25.658 ± 0.125	0.7478 ± 0.0047	0.1604 ± 0.0031
Obama	N	25.819 ± 0.062	0.7531 ± 0.0006	0.1580 ± 0.0013
Kjell	F	26.709 ± 0.563	0.7385 ± 0.0264	0.0836 ± 0.0028
Kjell	M	27.107 ± 0.143	0.7559 ± 0.0089	0.0798 ± 0.0018
Kjell	L	25.884 ± 0.037	0.7066 ± 0.0052	0.0972 ± 0.0046
Kjell	N	26.405 ± 0.103	0.7321 ± 0.0024	0.1001 ± 0.0015

A.2.3 Aggregation and resampling

Paired standard deviations are computed from the three Full–No-SACA seed differences within an identity. Macro-averages first average seeds within identity and then weight identities equally.

The interaction analysis follows the 20,000-replicate hierarchical bootstrap in Section 4.1, preserving four-configuration pairing. Effects retain original metric directions: negative LSE-D and positive LSE-C changes are favorable. Identity-specific and leave-out results accompany these exploratory three-identity intervals.

A.3 Seed-1 paired outcomes for all eighteen identities

All differences are Full–No-SACA at seed 1, with lip and mask supervision enabled. The four groups contain the core three identities, three HDTF-derived identities, two exploratory public-video identities, and ten further public-video identities. Direction counts use unrounded values.

A.3.1 Synchronization and PSNR changes

Table 9 lists every evaluated identity; Table 4 aggregates these same pairs. Each added identity has 273 whole-image evaluation frames. Coverage of mouth-frame records is documented in Appendix B.3.

Table 9: All 18 seed-1 Full–No-SACA pairs with lip and mask supervision. D/C: LSE-D/LSE-C; M-PSNR: mouth PSNR (dB). Better/Worse means both synchronization metrics improve/worsen; Mixed means discordant directions.

Identity	Group	Frames	ΔD	ΔC	ΔPSNR	ΔM-PSNR	Outcome
May	Core 3	553	-2.369	+2.761	-0.127	+0.285	Better
Obama	Core 3	728	+0.055	-0.029	-0.167	-0.149	Worse

Continued on next page

Table 11: Cross-source audio driving of known targets. All six directions use Full at seed 1.

Audio source → target	Sampled	Valid	ArcFace cosine
May → Kjell	80	79	0.933
Obama → Kjell	80	79	0.936
Kjell → May	80	80	0.910
Obama → May	93	93	0.911
Kjell → Obama	80	80	0.947
May → Obama	93	93	0.945

Table 9 continued

Identity	Group	Frames	ΔD	ΔC	ΔPSNR	ΔM-PSNR	Outcome
Kjell	Core 3	238	+0.005	+0.216	-0.066	-0.491	Mixed
Deb Haaland	HDTF 3	273	-1.253	+1.486	+0.037	-0.311	Better
Andy Kim	HDTF 3	273	+0.476	-0.651	-0.236	-0.249	Worse
Lauren Underwood	HDTF 3	273	+1.272	-1.096	+0.425	-0.246	Worse
Antony Blinken	Public 2	273	-1.464	+1.484	-0.307	-0.149	Better
Kamala Harris	Public 2	273	+0.540	-0.454	-0.699	-0.207	Worse
Joe Biden	Public 10	273	-0.188	+0.390	-1.260	-0.209	Better
Donald Trump	Public 10	273	-0.572	+0.718	-0.096	-0.842	Better
Michelle Obama	Public 10	273	-0.598	+0.682	-0.583	-0.158	Better
Melania Trump	Public 10	273	+0.246	+0.215	-0.030	-0.486	Mixed
Mike Pompeo	Public 10	273	-0.050	+0.215	-0.086	-0.276	Better
Anthony Fauci	Public 10	273	+0.071	+0.283	-0.653	-0.190	Mixed
Mike Pence	Public 10	273	+0.358	-0.607	+0.144	-0.430	Worse
Kayleigh McEnany	Public 10	273	-0.288	+0.362	-0.017	-0.441	Better
Karine Jean-Pierre	Public 10	273	-0.129	+0.355	+0.060	-0.077	Better
John Kirby	Public 10	273	-0.468	+0.584	-0.175	-0.341	Better

A.3.2 SSIM AND LPIPS CHANGES

Table 10: Whole-image and mouth (M-) structural/perceptual changes for all 18 seed-1 pairs. Positive SSIM and negative LPIPS changes are favorable. Signed 0.0000 values retain the direction of small changes before rounding.

Identity	ΔSSIM	ΔLPIPS	ΔM-SSIM	ΔM-LPIPS
May	-0.0034	+0.0033	+0.0053	+0.0005
Obama	-0.0002	+0.0005	-0.0042	+0.0046
Kjell	-0.0004	+0.0001	-0.0214	+0.0013
Deb Haaland	+0.0011	-0.0000	-0.0004	+0.0021
Andy Kim	-0.0003	+0.0012	+0.0013	+0.0089
Lauren Underwood	+0.0079	-0.0003	+0.0022	+0.0049
Antony Blinken	-0.0020	+0.0005	+0.0053	+0.0041
Kamala Harris	-0.0049	+0.0041	-0.0028	+0.0015
Joe Biden	-0.0014	+0.0194	-0.0194	+0.0038
Donald Trump	-0.0005	+0.0006	-0.0332	+0.0121
Michelle Obama	-0.0040	+0.0024	+0.0010	+0.0138
Melania Trump	+0.0004	-0.0004	-0.0045	+0.0028
Mike Pompeo	-0.0000	+0.0011	-0.0131	+0.0013
Anthony Fauci	-0.0029	+0.0003	-0.0039	+0.0056
Mike Pence	+0.0004	-0.0019	-0.0217	+0.0214
Kayleigh McEnany	+0.0017	-0.0031	-0.0172	-0.0002
Karine Jean-Pierre	+0.0007	+0.0000	-0.0004	-0.0009
John Kirby	-0.0010	+0.0007	-0.0168	+0.0027

A.4 ADDITIONAL CROSS-SOURCE EVALUATION DETAILS

Each direction uses the target identity's trained Full/seed-1 model. Generated and target-reference frames are paired by index. The InsightFace `buffalo_l` pipeline selects the largest detected

face in each image and computes the dot product of L2-normalized embeddings. Uniform-stride sampling retains at most 100 frames per direction. Pairs enter the average only when both faces are detected; Table 11 reports sampled and valid counts.

The evaluation describes known-target identity similarity under changed driving audio. Its scope and the absence of paired cross-source baselines are stated in Section 4.6.

A.5 Sensitivity to identity composition

Using the unrounded metrics for the core 36 runs, we calculate δ_{on}, δ_{off}, and $\delta_{\mathrm{interaction}}$ within each identity–seed pair, then average the three seeds. Table 12 reports identity-specific effects, and Table 13 recomputes their averages after omitting each identity.

Table 12: Identity-specific conditional effects and interactions, averaged over three paired seeds. D/C: LSE-D/LSE-C.

Identity	D δ_{on}	D δ_{off}	D $\delta_{\mathrm{interaction}}$	C δ_{on}	C δ_{off}	C $\delta_{\mathrm{interaction}}$
May	-1.247	+0.030	-1.277	+1.616	-0.117	+1.733
Obama	-0.145	+0.034	-0.179	+0.449	-0.158	+0.607
Kjell	+0.158	+0.838	-0.680	+0.167	+0.047	+0.120

Table 13: Leave-one-identity-out sensitivity. Each row reuses two core identities and all three seeds.

Omitted	Retained	D δ_{on}	C δ_{on}	D $\delta_{\mathrm{interaction}}$	C $\delta_{\mathrm{interaction}}$
May	Obama, Kjell	+0.006	+0.308	-0.430	+0.364
Obama	May, Kjell	-0.544	+0.892	-0.978	+0.927
Kjell	May, Obama	-0.696	+1.032	-0.728	+1.170

Each leave-out combination reuses two identities and their three seeds, providing a sensitivity summary of the same experiment. After excluding May, lip-on LSE-D has $\delta_{\mathrm{on}} = +0.006$ while its interaction is -0.430: the favorable interaction coexists with a slightly adverse conditional distance effect.

A.6 All core training runs

Tables 14 and 15 enumerate all 36 core runs, identified by identity, seed, and configuration. Aggregates use the unrounded records. Display precision is three decimals for PSNR and synchronization and four for SSIM/LPIPS.

Table 14: All 36 core runs: whole-image and synchronization metrics. F: Full; M: No-SACA; L: No-Lip; N: No-SACA-No-Lip. PSNR is in dB.

Identity	Seed	Config.	PSNR ↑	SSIM ↑	LPIPS ↓	LSE-D ↓	LSE-C ↑
May	1	F	28.828	0.8572	0.1009	8.699	4.717
May	1	M	28.955	0.8606	0.0976	11.068	1.956
May	1	L	28.536	0.8523	0.0965	11.671	1.105
May	1	N	28.818	0.8575	0.0957	11.591	1.195
May	2	F	28.820	0.8577	0.1001	9.658	3.593
May	2	M	29.166	0.8645	0.0976	9.899	3.049
May	2	L	28.457	0.8521	0.0970	11.550	1.138
May	2	N	28.917	0.8621	0.0959	11.658	1.301
May	3	F	28.887	0.8601	0.1041	9.346	3.862
May	3	M	29.081	0.8635	0.0966	10.475	2.318
May	3	L	28.674	0.8572	0.0978	11.806	0.937
May	3	N	28.998	0.8622	0.0927	11.687	1.033

Continued on next page

Table 14 continued

Identity	Seed	Config.	PSNR ↑	SSIM ↑	LPIPS ↓	LSE-D ↓	LSE-C ↑
Obama	1	F	35.225	0.9659	0.0269	10.313	2.950
Obama	1	M	35.391	0.9661	0.0263	10.258	2.978
Obama	1	L	35.007	0.9648	0.0267	11.497	0.773
Obama	1	N	35.393	0.9664	0.0260	11.694	0.804
Obama	2	F	35.173	0.9660	0.0272	10.237	3.590
Obama	2	M	35.394	0.9665	0.0269	10.471	2.640
Obama	2	L	35.016	0.9651	0.0269	11.671	0.718
Obama	2	N	35.347	0.9663	0.0263	11.686	0.815
Obama	3	F	35.355	0.9662	0.0269	9.857	3.918
Obama	3	M	35.607	0.9669	0.0260	10.115	3.493
Obama	3	L	34.881	0.9641	0.0276	11.825	0.636
Obama	3	N	35.325	0.9658	0.0258	11.511	0.982
Kjell	1	F	36.475	0.9833	0.0150	14.053	1.095
Kjell	1	M	36.542	0.9837	0.0149	14.048	0.879
Kjell	1	L	36.210	0.9827	0.0147	13.305	0.684
Kjell	1	N	36.418	0.9834	0.0146	12.297	0.789
Kjell	2	F	36.598	0.9837	0.0151	14.375	0.823
Kjell	2	M	36.626	0.9839	0.0148	14.309	0.641
Kjell	2	L	36.163	0.9825	0.0148	13.599	0.763
Kjell	2	N	36.385	0.9835	0.0146	12.668	0.578
Kjell	3	F	36.403	0.9832	0.0151	14.476	0.799
Kjell	3	M	36.629	0.9840	0.0150	14.072	0.696
Kjell	3	L	36.250	0.9828	0.0149	13.497	0.664
Kjell	3	N	36.427	0.9835	0.0147	12.923	0.603

Table 15: All 36 core runs: mouth reconstruction and valid mouth-frame counts. Configuration codes follow Table 14. M-PSNR is in dB.

Identity	Seed	Config.	Valid frames	M-PSNR ↑	M-SSIM ↑	M-LPIPS ↓
May	1	F	553	23.797	0.7547	0.1524
May	1	M	553	23.512	0.7494	0.1519
May	1	L	553	22.981	0.7383	0.1449
May	1	N	553	23.036	0.7405	0.1465
May	2	F	553	23.615	0.7548	0.1494
May	2	M	553	23.819	0.7598	0.1447
May	2	L	553	22.816	0.7366	0.1448
May	2	N	553	23.012	0.7431	0.1533
May	3	F	553	23.625	0.7533	0.1550
May	3	M	553	23.647	0.7555	0.1455
May	3	L	553	22.896	0.7386	0.1446
May	3	N	553	23.059	0.7428	0.1470
Obama	1	F	728	26.123	0.7608	0.1576
Obama	1	M	728	26.272	0.7650	0.1529
Obama	1	L	728	25.799	0.7528	0.1625
Obama	1	N	728	25.769	0.7526	0.1572
Obama	2	F	728	26.138	0.7625	0.1528
Obama	2	M	728	26.271	0.7643	0.1495
Obama	2	L	728	25.614	0.7469	0.1569
Obama	2	N	728	25.801	0.7537	0.1573
Obama	3	F	728	26.419	0.7669	0.1592
Obama	3	M	728	26.504	0.7700	0.1560
Obama	3	L	728	25.562	0.7436	0.1619
Obama	3	N	728	25.889	0.7531	0.1595
Kjell	1	F	236	26.464	0.7244	0.0823
Kjell	1	M	236	26.955	0.7458	0.0810
Kjell	1	L	236	25.845	0.7052	0.0957
Kjell	1	N	236	26.490	0.7331	0.0996
Kjell	2	F	236	27.352	0.7690	0.0817

Continued on next page

Table 15 continued

Identity	Seed	Config.	Valid frames	M-PSNR ↑	M-SSIM ↑	M-LPIPS ↓
Kjell	2	M	236	27.238	0.7624	0.0777
Kjell	2	L	236	25.888	0.7022	0.0936
Kjell	2	N	236	26.435	0.7338	0.0990
Kjell	3	F	236	26.310	0.7221	0.0868
Kjell	3	M	236	27.127	0.7596	0.0808
Kjell	3	L	236	25.919	0.7123	0.1023
Kjell	3	N	236	26.291	0.7294	0.1018

After excluding failed mouth detections, valid counts are May 553, Obama 728, and Kjell 236; Kjell has 238 whole-image frames.

B Configuration records and reproducibility scope

B.1 Architecture and training configuration

Table 16: Recorded architecture and training configuration.

Item	Recorded setting
Spatial planes	XY, YZ, XZ; base resolution 64; scale factors $\{1, 2\}$; 32 plane feature channels
Spatial/condition dimension	64; four-condition matrix 4×64
Audio input	1024-dimensional speech features; AudioNet, AudioAttNet, and mapping produce one frame-level token
Eye input	32-dimensional sine/cosine encoding, mapped to 64 dimensions
Camera input	12-dimensional input, mapped to 64 dimensions
SACA	One layer, two heads; feed-forward hidden width 128; dropout 0.2
Residual heads	Two linear layers with ReLU; hidden width 128; output dimensions 3, 3, 4, 1, 48
Training schedule	8,000 global iterations and 2,000 refinement iterations; batch sizes 1 and 32
Global loss weights	Image L_1: 0.8; VGG perceptual term: 0.01; $1 - \mathrm{SSIM}$: 0.2
Local fine-stage weights	Lip: 0.4; mask: 0.4. Lip-off configurations disable only the lip term
Deformation learning rate	Initial 0.0001; final 0.00001
Core seeds	1, 2, 3, paired within identity and configuration
Added identities	Full and No-SACA, both at seed 1

Table 17: Five independent Gaussian attribute residual heads.

Attribute	Residual	Output dimension
Position	$\Delta x_{i,t}$	3
Scale	$\Delta s_{i,t}$	3
Quaternion rotation	$\Delta r_{i,t}$	4
Opacity	$\Delta \alpha_{i,t}$	1
Spherical harmonics	$\Delta \mathrm{SH}_{i,t}$	16×3

Rotation uses quaternion parameters. The spherical-harmonic head outputs 48 components reshaped to 16×3. Tables 16 and 17 specify the recorded settings and interfaces; Appendix B.3 defines the scope of the reproducibility materials.

B.2 Comparison conditions and evaluation records

Table 18: Recorded external comparison conditions. Step counts and batches retain each method's own definition.

Configuration	Archived training record	Evaluation frames/seed
Full	8,000 global + 2,000 refinement iterations; batches 1/32	May 548, Obama 724, Kjell 238; seed 1
GaussianTalker	10,000-step checkpoint; batch 128	Same-identity common validation prefix; seed 1
GeneFace++	Head + torso; recorded as 250,000 steps	Same-identity common validation prefix; seed 1
ER-NeRF	Head + lip + torso; complete step counts not uniformly listed	Same-identity common validation prefix; seed 1

Training steps and batch sizes follow each method's own definition. The archived systems use method-specific compute budgets, background and torso processing, checkpoint selection, and preprocessing, so the aligned-frame results are interpreted as a system-level comparison. Table 18 records these conditions.

Core factorial evaluation uses May 553, Obama 728, and Kjell 238 whole-image frames, separately from the external prefixes. Reconstruction metrics average valid frames within each model. Mouth crops use reference landmarks and a $1.5\times$ bounding box. Synchronization is evaluated on generated clips using 25 fps video and 16 kHz mono audio. Identity-similarity sampling and detection exclusions are detailed in Appendix A.4.

B.3 Scope of the reproducibility materials

The supplementary archive contains experimental configurations, analysis and evaluation scripts, per-configuration results, and provenance records. Together with Sections 3.2–3.5 and Tables 16–17, these materials support regeneration of the reported tables, plots, descriptive analyses, and sensitivity summaries from recorded outputs.

The release is designed for analysis-level reproduction. End-to-end retraining and rendering additionally depend on the complete GaussianTalker-based implementation and its original preprocessing environment. Implementation-dependent details include spatial-plane sampling, attention and residual ordering, rendering-attribute conversion, dataset split identifiers, mask and lip-crop preprocessing, and software and scorer versions. This scope distinguishes results that can be regenerated from the supplementary archive from full system reconstruction.

B.4 Inherited design and the present study

Table 19 attributes the base spatial-query deformation architecture to GaussianTalker (Cho et al., 2024) and identifies the configuration choices and empirical analyses contributed by this study.

Table 19: Base architecture, recorded configuration, and this study's empirical contributions.

Component	Origin/nature	Treatment in this study
Multi-resolution tri-planes and Gaussian-wise spatial features	GaussianTalker representation design	Retain the encoder interface; report resolution, scales, and dimensions

Continued on next page

Table 19 continued

Component	Origin/nature	Treatment in this study
SACA, four conditions, and five attribute residuals	GaussianTalker conditional deformation	Credit the architecture; specify condition encoding, fusion interface, and residual settings
Staged optimization and lip reconstruction	Training strategy used by GaussianTalker	Report the schedule, objectives, and weights in the current experiments
Foreground-mask term	Archived training-script configuration	Preserve the recorded transform, weight, and L_1 definition; treat target preprocessing as implementation-dependent
Four-condition mean replacement crossed with lip on/off	Factorial comparison in this study	Define conditional configuration effects and interaction; document active-capacity and path differences
Three identities/three seeds, fifteen added identities, six cross-source directions	Experiments and analyses in this study	Report interaction, identity variation, synchronization–reconstruction trade-offs, and identity similarity

C Figure provenance and qualitative selection

C.1 Interaction plots and scatter-plot conventions

Figure 2a uses the 36 rows labeled `original3` in `core_and_extension_66_cells.csv`, with seed-within-identity aggregation as specified in Section 4.1. Conditional effects and interactions use unrounded metrics. Table 3 retains the archived hierarchical bootstrap intervals. The original LSE-D archive used favorable benefit directions; conversion to raw distance transforms interval endpoints as $[-\text{upper}, -\text{lower}]$.

Figure 2b uses all eighteen rows of `all18_seed1_paired.csv`. Coordinates are unrounded Full–No-SACA differences: LSE-D horizontally and LSE-C vertically. Both metrics improve when $\Delta\text{LSE-D} < 0$ and $\Delta\text{LSE-C} > 0$ and worsen with the opposite signs; discordant directions are labeled mixed. The upper-left, upper-right, lower-right, and lower-left quadrants contain 10, 3, 5, and 0 identities, respectively. Classification uses effect direction, with no significance test applied to individual points.

Color represents mouth PSNR change on a zero-centered $[-0.9, +0.9]$ dB range: brown is negative and blue positive. All values fall within this range. Numeric labels follow source-table order. Points use unjittered coordinates; leader lines and Appendix A.3 identify nearby or overlapping points.

C.2 Experimental images and mouth enlargements

Figure 3 uses archived montages under `02_source_results`. The identity, relative file, and selected archive column are:

- May: `04_six_scene_reanalysis/results/qualitative/may_gt_full_no_saca.png`, column 3.
- Melania Trump: `06_public_video10/results/qualitative/melaniatrump_gt_full_no_saca.png`, column 3.
- Lauren Underwood: `03_public3/results/qualitative/laurenunderwood_gt_full_no_saca.png`, column 1.

Source rows are Ground truth, Full, and No-SACA, reordered as Reference, No-SACA, and Full in the figure and enlargements. Within an identity, all images use the same archived column. The montage script sorts the intersection of common frame keys and selects five evenly spaced positions. Column labels identify the archive positions used for the qualitative comparison.

All cases use seed-1 test outputs at 10,000 iterations. May source thumbnails are 192×192 pixels with 28-pixel row headers; the others are 256×256 with 34-pixel headers. Identities were selected to represent the three synchronization effect directions, as described in Section 4.7. This direction-based sample supports qualitative case analysis; estimating failure rates would require representative sampling.

With the upper-left source-thumbnail corner as the origin, mouth windows use half-open coordinates $[\text{left}, \text{top}, \text{right}, \text{bottom})$: May $[57, 110, 125, 154)$, Melania Trump $[95, 106, 163, 150)$, and Lauren Underwood $[102, 150, 170, 194)$. Reference, No-SACA, and Full use the same 68×44 window within identity and preserve its aspect ratio. These coordinates define both the boxes on full thumbnails and the enlarged crops. They are presentation windows, distinct from the landmark-based crops used for mouth metrics.

Processing is limited to cropping, reordering, and proportional display of the archived thumbnails. Their source resolution determines the available detail.

C.3 Framework configuration correspondence

The diagram follows the recorded batch interface: temporal audio is aggregated before four-condition fusion, two-scale spatial features supply the queries, and five heads predict Gaussian attribute residuals. One of the two SACA heads is illustrated. Reproducibility scope is documented in Appendix B.3.

Source-specific audio-gated dynamic planes, anisotropic attention, hash positional encoding, and HuBERT layer-weighted fusion are excluded from the studied configuration. The mean replacement and factorial design are specified in Section 3.5 and Table 1.