# CONFERENCING TOKENS: LEARNING COMPACT HUMAN REPRESENTATIONS FOR VIDEO COMMUNICATION

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Video conferencing serves both human participants and intelligent assistants, yet pixel-based communication requires models to decode video and process long visual input sequences. We introduce Conferencing Tokens, a compact, transmissible representation of people in motion that connects the same decoded output to different downstream models. Initialized from pretrained geometry, the tokenizer combines rate–distortion learning with causal temporal entropy modeling to form a shared framewise interface. The decoded representation supports video reconstruction with person-specific appearance models and enters a language model directly through a learned interface, using one input position per frame. Experiments show a mean BD-PSNR gain of 0.53 dB over a facial-parameter-based conferencing coding baseline. For transcript-conditioned meeting-state prediction, language adaptation retains accuracy close to an unadapted dense RGB reference while reducing modality input positions by a factor of 72.6. These results demonstrate the feasibility of reusing a compact communication representation directly for downstream analysis, linking efficient transmission with model consumption.

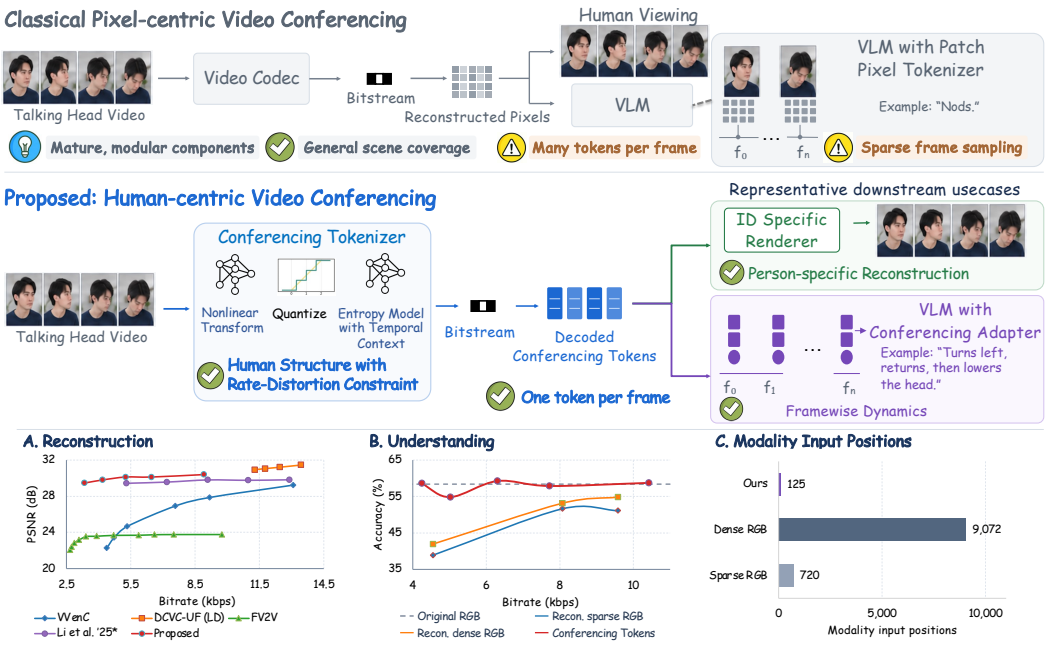


Figure 1: Conferencing Tokens connect a compressed human representation to person-specific video reconstruction and language-based behavior understanding. The pixel pipeline transmits frames and then tokenizes them for a VLM; our pipeline provides one Conferencing Token input position per frame to a learned language interface. Lower panels summarize reconstruction quality, accuracy on the understanding task, and modality input-position counts.

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.

# 1 INTRODUCTION

Video conferencing is becoming a shared environment for human participants and intelligent assistants. Beyond speech, facial expressions, head movements, and head orientation convey nonverbal cues that help people follow interactions and provide context for automated meeting analysis. Recent meeting-understanding benchmarks examine the integration of verbal and nonverbal evidence (Wang et al., 2026). The same evolving visual information therefore serves two consumers: remote participants viewing one another and meeting assistants analyzing their interactions.

Pixels provide a general, modular interface between communication and these consumers, with video codecs efficiently exploiting spatial and temporal redundancy. For language-based analysis, the received stream is decoded into frames and then visually encoded into model inputs. Dense inputs can incur substantial token costs, while sparse sampling under a fixed budget reduces temporal coverage. This general visual interface is not explicitly organized around the person and their behavior, motivating a more direct, compact interface based on the information those consumers need.

Prior work on human–machine and multipurpose compression connects reconstruction with machine analysis (Choi & Bajić, 2022a;b), while learned visual interfaces connect visual representations to language models (Li et al., 2023). These precedents raise a specific question for conferencing: can we learn a compact, transmissible representation of changing human behavior whose decoded form both preserves reconstructability and directly supports downstream models? Addressing this question calls for jointly evaluating transmission bitrate, reconstruction quality, understanding performance, and language-model input length.

We introduce *Conferencing Tokens*, a shared dynamic representation extracted from person video, transmitted as a bitstream, and reconstructed at the receiver (Figure 1). Geometric initialization, geometric constraints, and causal temporal entropy models guide the learning of a compressible representation of facial motion, head pose, and root motion. The same decoded representation drives person-video reconstruction with a prebuilt person-specific appearance model and directly enters a language model through a learned language interface. The tokenizer remains fixed during downstream adaptation.

We evaluate person-video reconstruction and meeting-state prediction with shared transcripts. Reconstruction achieves a mean BD-PSNR gain of $+0.53$ dB over a facial-parameter-based conferencing codec baseline (Li et al., 2025). After language adaptation, the interface achieves overall accuracy close to unadapted dense RGB on the evaluated understanding task, while reducing modality input positions from 9,072 to 125 per five-second example.

Our contributions are summarized as follows:

1. **A shared representation for communication and downstream models.** Conferencing Tokens connect transmission with person-video reconstruction and language analysis through the same decoded representation.
2. **Representation learning for efficient transmission.** Geometric priors, learned latent and physical parameter branches, and causal temporal entropy models preserve human structure and motion under a rate constraint.
3. **Evidence across reconstruction and understanding.** Experiments demonstrate favorable reconstruction rate–distortion performance and show that language adaptation retains meeting-state prediction accuracy with compact model inputs.

# 2 RELATED WORK

**Human Representations for Transmission.** Generative face video coding separates changing facial motion from appearance, using compact keypoints in Face-vid2vid (Wang et al., 2021) or learned facial features in CFTE (Chen et al., 2022) to drive reconstruction. Geometry-based avatars provide an explicit reconstruction interface: FlashAvatar deforms Gaussians attached to a parametric face model (Xiang et al., 2024), while the conferencing codec of Li et al. (2025) transmits facial parameters to drive a Gaussian head model. Human–machine coding also connects reconstruction and analysis through scalable streams or human-centric representations (Choi & Bajić, 2022b; Jiang

et al., 2023b). Following the separation of motion and appearance, we learn a decoded human interface under a rate constraint and investigate its use beyond pixel reconstruction.

**Structured Human Representations for Language Models.** Structured human representations offer a language-model interface for motion and facial information. MotionGPT models tokenized body motion and text in a unified language framework (Jiang et al., 2023a). UniTalker uses a neural landmark codec to incorporate facial expression sequences into multimodal dialogue and audiovisual response generation (Hu et al., 2025); TDMM-LM connects 3D facial parameters with language for motion description and text-conditioned facial animation (Song et al., 2026). We study human representations learned under transmission-rate constraints, adapting the decoded outputs of a fixed tokenizer to reconstruction and language interfaces.

**Compact Multimodal Language Interfaces.** Compact multimodal interfaces address visual sequence length and the connection between representation spaces. LLaVA-Mini combines modality pre-fusion with visual compression before the language model (Zhang et al., 2025); FastVID reduces visual sequences through temporal segmentation and token pruning (Shen et al., 2025), and VoCo-LLaMA learns compression tokens through the language model (Ye et al., 2025). BLIP-2 connects frozen pretrained visual and language models through a learned interface (Li et al., 2023). Our language interface receives a human representation already decoded from the communication stream, reusing the transmitted information for downstream analysis. We therefore evaluate communication bitrate and language-model input length separately, reflecting transmission and model-consumption costs.

# 3 METHOD

## 3.1 SYSTEM OVERVIEW AND PROBLEM FORMULATION

The communication representation must support quantization and entropy coding, preserve human geometry and motion under a rate constraint, and provide a shared set of decoded values that directly connects to downstream models.

For a person-video sequence $x_{1:T}$ of duration $\tau$, we encode facial motion, head pose, and root motion into a bitstream $b$. Decoding reconstructs a sequence of Conferencing Tokens $H = \{h_t\}_{t=1}^{T}$:

$$b = \mathcal{C}\big(Q(\mathcal{E}(x_{1:T}))\big), \qquad H = \mathcal{D}_{\mathrm{rec}}(b), \qquad h_t = (\hat{e}_t, \hat{p}_t), \tag{1}$$

where $\mathcal{E}$ is the human encoder, $Q$ quantization, $\mathcal{C}$ entropy coding including the required hyperprior symbols, and $\mathcal{D}_{\mathrm{rec}}$ entropy decoding and reconstruction of the numerical values. Each frame token combines a learned latent representation $\hat{e}_t$ and explicit physical controls $\hat{p}_t$; both downstream interfaces consume this same decoded pair. The dynamic transmission bitrate is $R = |b|/\tau$, where $|b|$ is the coded stream length in bits and $\tau$ is the sequence duration in seconds.

We instantiate two representative interfaces. A geometry decoder and head $\mathcal{G}$ map $H$ to geometry rendered with prebuilt person-specific appearance assets $A$. A language adapter $\mathcal{P}$ maps $H$ to input embeddings $Z$ for a pretrained language model $\mathcal{L}$:

$$\hat{x}_{1:T} = \mathcal{R}_A\big(\mathcal{G}(H)\big), \qquad Z = \mathcal{P}(H), \qquad \hat{y} = \mathcal{L}(Z, q, c), \tag{2}$$

Here, $q$ specifies the task and $c$ optional textual context. The language interface maps decoded token values directly to model input positions.

Training first learns the tokenizer through rate-constrained geometric reconstruction, then fixes it during downstream adaptation (Figure 2). Language adaptation changes how the model reads the representation while preserving the transmitted format and decoded values used for reconstruction.

## 3.2 RATE-CONSTRAINED TEMPORAL TOKENIZER

**Geometry-aware initialization.** We start from the pretrained image-to-geometry mapping of SHeaP (Schoneveld et al., 2025) and introduce an intermediate representation subject to quantization and rate constraints. A full-rank SVD factorizes the original local prediction heads into analysis and synthesis transforms, preserving their mapping before quantization. New residual branches

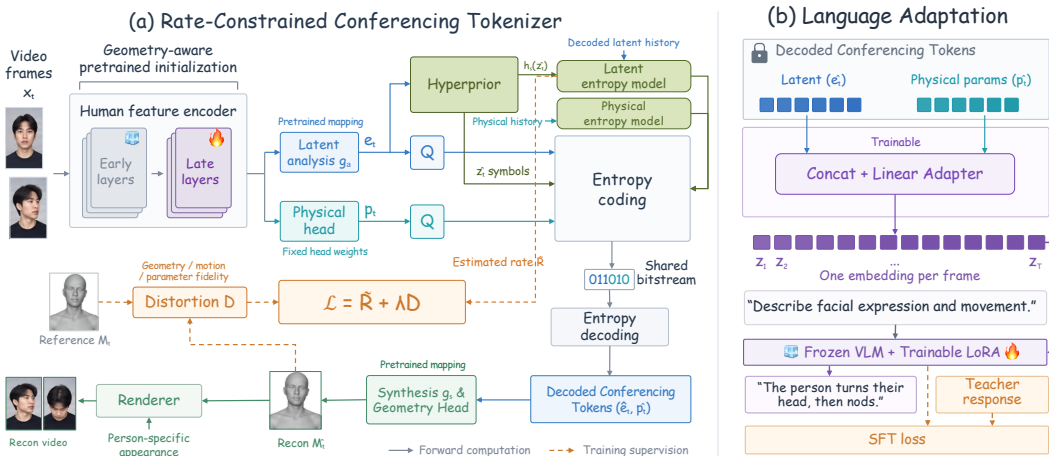


Figure 2: Two-stage learning. (a) The decoded tokens are shared by geometry reconstruction and language adaptation. The geometry decoder recovers person geometry for rendering, while the language adapter consumes the tokens directly. Person-specific appearance is outside the dynamic stream. Dashed arrows indicate supervision. (b) With tokenizer/base-model weights fixed, SFT trains one linear adapter and LoRA; the teacher is absent at inference.

have zero-output initialization. We fix the earlier visual layers and original physical-head weights while optimizing later visual layers, analysis/synthesis transforms, and entropy models, allowing the features and intermediate representation to adapt under geometric supervision. Appendix A.1.2 specifies the initialization and training configuration.

**Latent representation and physical parameters.** The latent branch encodes local deformation through a learned intermediate representation; synthesis recovers expression, eye-rotation, and eyelid controls. The physical branch retains explicit coordinates for pose and placement, with separate quantization. For frame features $f_t$, analysis produces $e_t = g_a(f_t) \in \mathbb{R}^{d_e}$ with $d_e = 108$, while the physical head produces $p_t \in \mathbb{R}^{d_p}$ with $d_p = 12$, comprising root rotation, translation, neck rotation, and jaw rotation. The branches are separately quantized and entropy modeled under joint geometric reconstruction supervision. Their decoded values form $h_t = (\hat{e}_t, \hat{p}_t)$: reconstruction applies synthesis to $\hat{e}_t$, whereas the language interface consumes the concatenated pair before synthesis. Parameter groups and quantization rules are given in Appendix A.1.2.

**Temporal entropy models.** Visual encoding remains framewise; causal Transformer entropy models use decoded history to predict current symbol probabilities, exploiting temporal correlations while retaining the framewise interface. Following hyperprior coding (Ballé et al., 2018), the latent branch derives $z_t = h_a(e_t) \in \mathbb{R}^{d_z}$ and quantizes it. Hyperlatent symbols are decoded first; their synthesized information $h_s(\hat{z}_t)$ combines with reconstructed latent history $\hat{e}_{<t}$ to predict the latent-symbol mean and scale. The physical model conditions only on reconstructed physical history $\hat{p}_{<t}$, without a separate hyperprior. Both models use receiver-reconstructable history, providing consistent conditions for encoding and decoding. The estimated code length in bits per frame is

$$\widetilde{R} = \frac{1}{T}\sum_t \big[ -\log_2 P_z(\hat{z}_t) - \log_2 P_e(\hat{e}_t \mid h_s(\hat{z}_t), \hat{e}_{<t}) \\ - \log_2 P_p(\hat{p}_t \mid \hat{p}_{<t}) \big]. \tag{3}$$

Here, each $P$ is the discrete probability mass of a quantized symbol under its logistic model, and hatted variables denote reconstructed values. Hyperlatents recover coding conditions without adding language-model input positions. History windows, resets, and decoding details appear in Appendix A.1.2.

**Rate–distortion training and measured rate.** We jointly optimize the representation and entropy models, balancing estimated transmission cost with geometric and motion fidelity. With forward-

pass rounding and straight-through gradients, the objective is

$$\mathcal{L}_{\text{tok}} = \widetilde{R} + \lambda D, \qquad \lambda > 0, \tag{4}$$

where $D$ compares decoded geometry and controls with references from the pretrained geometric model. Local geometric distortion constrains deformation; world-space distortion constrains geometry after pose and translation. Local velocity matches changes in deformation between adjacent frames, while non-expression parameter fidelity directly constrains rotations, translation, and other explicit controls alongside vertex supervision. Tokenizer learning uses these geometry and motion objectives; person-specific appearance reconstruction is handled by downstream renderer adaptation. Loss weights and normalization are specified in Appendix A.1.2.

Evaluation measures actual dynamic payload bitrate by entropy coding the symbols of both branches and the latent-branch hyperlatents, with the counting convention in Appendix A.1.2.

### 3.3 Reconstruction for Human Viewing

For reconstruction in Equation 2, synthesis maps $\hat{e}_t$ to local FLAME coordinates (Li et al., 2017), combined with physical parameters $\hat{p}_t$ and static shape $s$. FlashAvatar (Xiang et al., 2024) deforms person-specific Gaussians using this geometry. Prebuilt shape and appearance supply static identity information and visual detail outside the dynamic stream. The renderer consumes the decoded representation. Root motion has no independent shoulder/arm/torso joints; "upper body" denotes the evaluation region.

### 3.4 Language Adaptation Interface

**Continuous input interface.** Entropy decoding reconstructs numerical vectors $(\hat{e}_t, \hat{p}_t)$ from transmitted symbols. An affine adapter $\mathcal{P}_\phi : \mathbb{R}^{d_e+d_p} \to \mathbb{R}^{d_{\text{LM}}}$ maps their concatenation into the language model's continuous embedding space of dimension $d_{\text{LM}}$. The resulting embeddings serve as soft inputs rather than language-vocabulary token identifiers. This avoids text serialization or additional vocabulary-based quantization while connecting the communication representation to the model's input space.

**Framewise organization.** Each frame occupies one input position, retaining framewise motion information without temporal pooling or subsampling. A learned embedding preserves the position of a missing or invalid frame:

$$\zeta_t = \begin{cases} W_\phi[\hat{e}_t; \hat{p}_t] + b_\phi, & \text{if frame } t \text{ is valid,} \\ \zeta_{\text{miss}}, & \text{otherwise.} \end{cases} \tag{5}$$

The trainable parameters are $W_\phi$, $b_\phi$, and $\zeta_{\text{miss}}$. The sequence $Z = \{\zeta_t\}_{t=1}^{T}$ accompanies $q$ and $c$; the language model handles interactions across frames. Input organization and dimensions appear in Appendices A.3 and A.1.3.

**Answer-supervised adaptation.** Attention LoRA (Hu et al., 2022) adapts how the model reads the projected inputs and generates answers. We jointly optimize it with the adapter, freezing the tokenizer and base model. Given fixed responses $y^*$, answer-only SFT minimizes

$$\mathcal{L}_{\text{SFT}} = -\frac{1}{|\mathcal{A}|} \sum_{i \in \mathcal{A}} \log p_{\phi,\Delta}(y_i^* \mid \mathcal{P}_\phi(H), q, c, y_{<i}^*), \tag{6}$$

where $\mathcal{A}$ indexes answer positions including the end-of-sequence token, $\phi$ denotes the adapter parameters, and $\Delta$ the attention LoRA parameters. Prompt and human-input positions are excluded from the loss. The targets combine dense-RGB teacher responses with agent replacements (Appendix A.1.1). Exploratory response-distillation and hidden-supervision objectives are reported separately in Appendix A.4.

# 4 EXPERIMENTS

## 4.1 EXPERIMENTAL SETUP

**Training data and configuration.** Tokenizer data comprise 10,925 accepted clips from AMI (Carletta, 2006), CCDb+ (Deng et al., 2025), NeRSemble (Kirschstein et al., 2023), Open3DFaceVid (Song et al., 2026), and VFHQ (Xie et al., 2022). The same video pool supplies 12,015 training and 1,411 validation question–answer pairs. Five target rates (3/4.5/6/7.5/9 kbps) have separate downstream adaptations; we report measured payload rates. Data and training details appear in Appendix A.1.

**Reconstruction evaluation.** Eight NeRFBlendShape (Gao et al., 2022) identities use $512 \times 512$ frames: the final 500 per identity are tested; the remainder train separate renderers per identity/rate. Training-subset PSNR selects one checkpoint for all metrics (Appendix A.2.2). Baselines are VVenC (Wieckowski et al., 2021)/VVC (Bross et al., 2021), low-delay DCVC-UF (Li et al., 2026), FV2V (Wang et al., 2021), and Li et al. '25* (Li et al., 2025), adapted to SHeaP inputs. Shared foreground masks define YUV420 PSNR, RGB SSIM (Wang et al., 2004), and LPIPS (Zhang et al., 2018) evaluation.

**Understanding evaluation.** STATE (Wang et al., 2026) predicts seven meeting-state labels from five-second clips and shared transcripts. Original RGB, VVenC-decoded RGB, and Conferencing Tokens share 589 paired examples. Proposed uses SFT; RGB references receive no corresponding adaptation. Metrics are accuracy, macro-F1, and modality input positions. Protocols and CREMA-D (Cao et al., 2014)/FEC-Bench (Zhao et al., 2026) subsets appear in Appendix A.3.

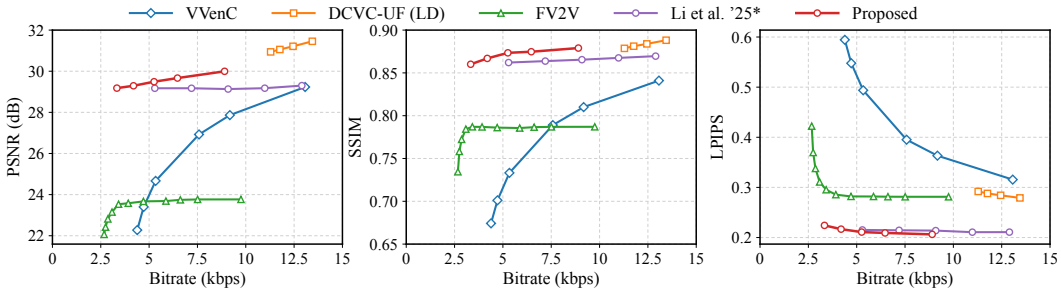


Figure 3: Reconstruction rate–distortion over eight identities, equally averaged. Foreground PSNR, SSIM, and LPIPS use common evaluation regions and measured dynamic bitrate. Li et al. '25* denotes our adapted implementation. Appendix A.2 specifies evaluation and static-asset accounting.

## 4.2 EVALUATION ON REPRESENTATIVE DOWNSTREAM TASKS

**Reconstruction rate–distortion.** Figure 3 shows competitive foreground reconstruction in the evaluated low-rate range. Relative to parameter-based conferencing coding, Conferencing Tokens achieve comparable or better quality at lower rates: PSNR and SSIM improve over much of the overlapping range, while LPIPS is close or modestly lower. The mean BD-PSNR gain over Li et al. '25* is $+0.53$ dB across eight identities (Appendix A.2.3). This comparison uses person-specific geometric priors and prebuilt appearance assets. The general video codecs cover different rate ranges; DCVC-UF (LD) achieves higher PSNR above 11 kbps.

**Understanding quality and input cost.** Table 1 shows that compact transmitted inputs retain STATE accuracy close to original RGB. At 6.30 kbps, the adapted Qwen3-VL-8B-Instruct (Bai et al., 2025) achieves 59.25% accuracy using 125 Conferencing Token input positions per five-second example, compared with 58.40% and 9,072 positions for dense RGB, a $72.6\times$ reduction in modality input length. Across the evaluated operating points, Proposed obtains 54.84–59.25% accuracy versus 38.88–54.84% for VVenC-decoded RGB, with fewer input positions; these are comparisons across rate ranges under the adaptation conditions in Section 4.1. At the representative point, macro-F1 is 21.91%, close to dense RGB's 22.38% and below sparse RGB's 26.19%. Class-level

results and cross-scenario evaluation appear in Appendices A.3.3 and A.3.4. Adapting coded facial parameters yields STATE performance similar to Conferencing Tokens, suggesting that the learned compact representation retains comparable task-relevant information (Appendix A.3.7).

To test dependence on the matched person representation, we fix the model, question, and transcript and replace the person-token sequence with one from another meeting. At 6.30 kbps, accuracy decreases from 59.25% to 40.75%, with decreases at all five operating points. This indicates that the adapted model uses person dynamics matched to the question and textual context (Appendix A.3.2).

Table 1: STATE prediction and input cost. Original RGB receives no additional coding. Local models use Qwen3-VL-8B. $N$ distinguishes the full 600-example set from the common 589-example subset used for Proposed and VVenC comparisons; published results are contextual references.

| | Model | $N$ | Rate (kbps)↓ | Acc. (%)↑ | Macro-F1 (%)↑ | Tokens↓ |
|---|---|---|---|---|---|---|
| Original RGB | Chance† | 600 | — | 14.29 | 14.29 | — |
| | Human† | 600 | — | 86.33 | 75.79 | — |
| | Gemini-3 Pro† | 600 | — | **59.00** | **30.39** | — |
| | GPT-4o† | 600 | — | 36.06 | 14.23 | — |
| | GPT-5† | 600 | — | <u>55.67</u> | <u>28.69</u> | — |
| | Qwen2.5-VL-32B† | 600 | — | 55.50 | 26.57 | — |
| | Qwen3-VL-8B† | 600 | — | 42.17 | 23.74 | — |
| | Qwen3-VL-32B† | 600 | — | 38.00 | 19.89 | — |
| | Qwen3-VL-8B (Sparse, 2 fps) | 600 | — | **59.00** | 26.32 | 720 |
| | Qwen3-VL-8B (Dense, 25 fps) | 600 | — | **59.00** | 22.51 | 9,072 |
| | Qwen3-VL-8B (Sparse, 2 fps) | 589 | — | 58.40 | 26.19 | 720 |
| | Qwen3-VL-8B (Dense, 25 fps) | 589 | — | 58.40 | 22.38 | 9,072 |
| VVenC-decoded RGB | Qwen3-VL-8B (QP63, 2fps) | 589 | 4.55 | 38.88 | 17.85 | 720 |
| | Qwen3-VL-8B (QP63, 25fps) | 589 | 4.55 | 41.94 | 20.51 | 9,072 |
| | Qwen3-VL-8B (QP52, 2fps) | 589 | 8.06 | 51.61 | <u>22.89</u> | 720 |
| | Qwen3-VL-8B (QP52, 25fps) | 589 | 8.06 | <u>53.14</u> | 22.55 | 9,072 |
| | Qwen3-VL-8B (QP50, 2fps) | 589 | 9.58 | 51.10 | 21.87 | 720 |
| | Qwen3-VL-8B (QP50, 25fps) | 589 | 9.58 | **54.84** | **23.23** | 9,072 |
| Conferencing Tokens | Proposed ($\lambda = 12.5$) | 589 | 4.24 | 58.57 | <u>22.35</u> | **125** |
| | Proposed ($\lambda = 814$) | 589 | 5.02 | 54.84 | 20.52 | **125** |
| | Proposed ($\lambda = 7687$) | 589 | 6.30 | **59.25** | 21.91 | **125** |
| | Proposed ($\lambda = 24651$) | 589 | 7.71 | 57.89 | 21.00 | **125** |
| | Proposed ($\lambda = 70000$) | 589 | 10.41 | <u>58.74</u> | **23.13** | **125** |

† Published 600-example MeetingToM results (Wang et al., 2026); paired local comparisons use $N = 589$. Bold/underline indicate best/second-best within each block; ties share a mark. Only the Human and Chance and the two $N = 589$ rows in the Original RGB block are excluded from that block's ranking. All rows in the VVenC-decoded RGB and Proposed blocks participate in their respective block rankings. Tokens are modality input positions, not coded symbols or total context length. $\lambda$ is the tokenizer rate–distortion training parameter. Sampling and rate accounting are detailed in Appendix A.3.

**Qualitative reconstruction and understanding.** Figure 4 shows that Conferencing Tokens retain recognizable eyelid closure and mouth deformation at low dynamic bitrates. The enlarged crops reveal differences in eyelid boundaries and mouth contours across reconstructions. For the corresponding two-second clips, the language model directly consumes 50 Conferencing Token input positions to describe blinking and mouth opening/closing. Dense RGB also identifies these actions, while sparse RGB answers *No change* in both examples. These examples illustrate how the same decoded representation preserves visible facial changes for viewing and supports action descriptions through a compact language interface.

### 4.3 Effect of Temporal Entropy Context

We test whether decoded history improves coding efficiency while preserving the framewise interface. At $\lambda = 7687$, No history, E history, P history, and E+P history retain the same network capacity and current-frame hyperprior; masking controls which histories are used. Figure 5 marks

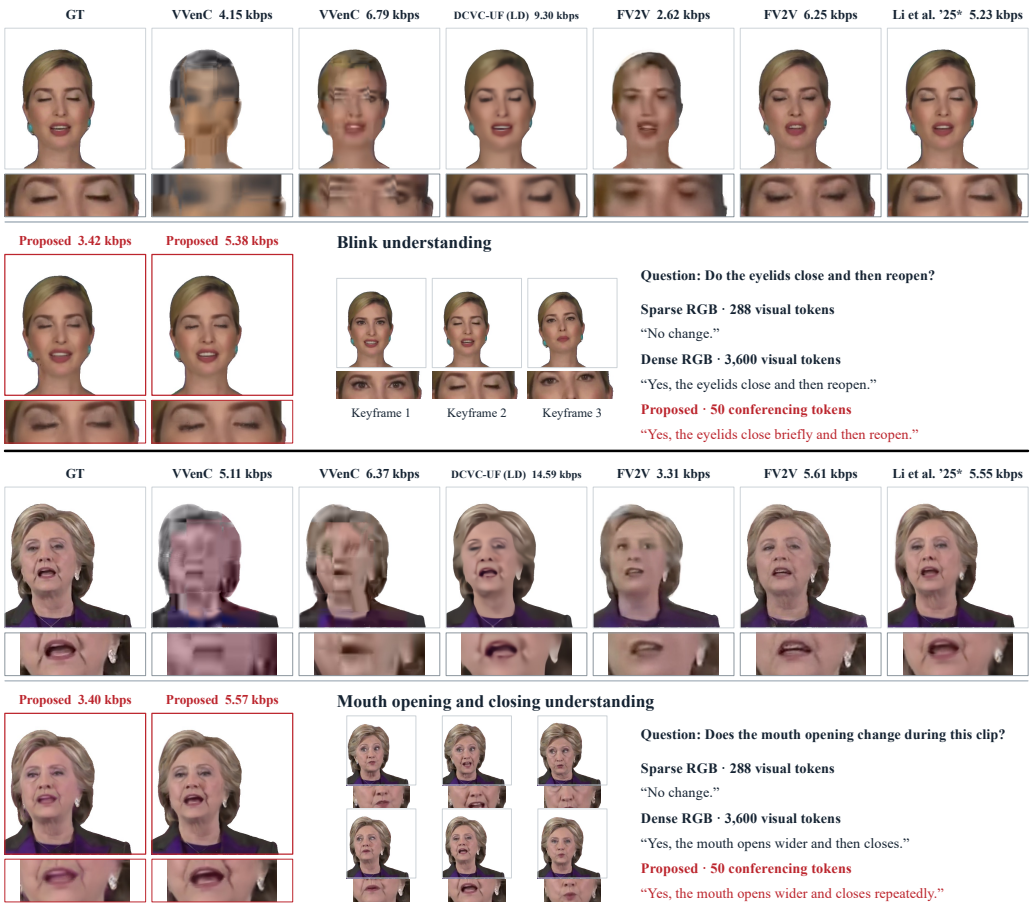


Figure 4: Reconstruction and understanding of a blink (top) and mouth opening/closing (bottom). Reconstructions show the same instant and local crops under a common foreground mask on white; rates are measured over each person's full 500-frame sequence and are not strictly matched. Understanding uses two-second clips: sparse/dense RGB take 4/50 VVenC-reconstructed frames, while Proposed uses 50 input positions. Keyframes illustrate the actions. Appendix A.3.6 gives input details.

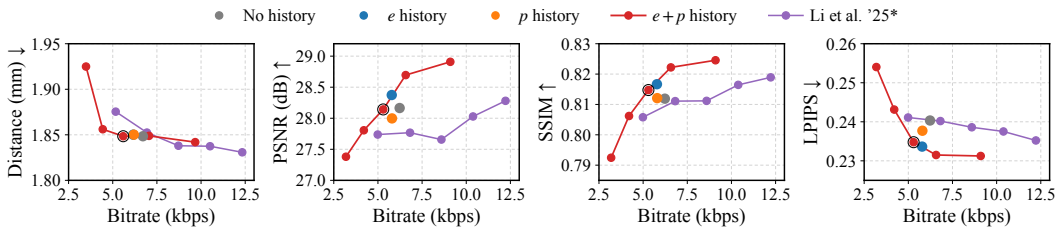


Figure 5: Temporal entropy context. Geometry averages ten identities; reconstruction metrics and rates equally average ID 1/ID 2. The three isolated configuration points and the black-ringed red E+P point form the $\lambda = 7687$ ablation. The full red E+P and purple Li et al. '25* D10–D50 curves provide rate–distortion context. Rates are measured; protocols appear in Appendix A.2.5.

the E+P ablation point with a black ring, distinguishing this comparison from the full background curves.

On the independent ten-person geometry set, average distance stays near 1.85 mm while E+P reduces bitrate from 6.71 to 5.59 kbps (16.6%). On the separate two-person reconstruction set, bitrate

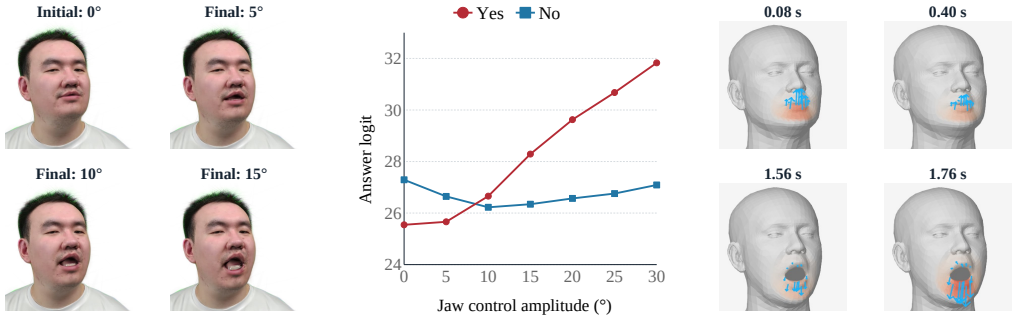


Figure 6: Controlled mouth-opening analysis. Left: rendered examples at selected amplitudes. Middle: Yes/No answer logits over the full $0°$–$30°$ sweep. Right: geometric sensitivity at selected times in the $15°$ sequence, favoring closing early and opening late. Colors show displacement magnitude; arrows indicate local directions that increase the Yes–No logit difference.

decreases from 6.23 to 5.30 kbps (14.9%), with nearly unchanged PSNR and modest improvements in SSIM and LPIPS. E-only yields the best reconstruction metrics among the four configurations, whereas E+P further reduces rate. Thus, causal history improves codability with broadly stable geometric and reconstruction quality in the evaluated settings (Appendix A.2.5).

### 4.4 Controlled Geometric Sensitivity of Language Predictions

We use controlled jaw motion to examine how language answers respond to geometric changes and interpret that response through the shared geometry decoder. For the 6 kbps-target SFT model, we vary jaw-opening amplitude while fixing the latent representation and other physical controls. Each two-second sequence receives the same question about whether the mouth becomes more open. The two answer-logit curves in Figure 6 show an increasing advantage of Yes over No as opening amplitude grows.

Let $s(H; q) = \ell_{\text{Yes}} - \ell_{\text{No}}$. We compute an input direction that increases this score, project it onto latent and jaw coordinates, and map it through the geometry decoder to vertex-displacement directions:

$$d_M = J_{\mathcal{G}}(H)\, P_{e,\text{jaw}} \nabla_H s(H; q). \tag{7}$$

In the $15°$ example, sensitivity concentrates around the mouth and jaw, favoring closing at selected early positions and opening at later positions. These directions accentuate the beginning-to-end opening contrast queried by the question. Arrows represent local answer-score sensitivity rather than the person's motion trajectory. The shared representation thus makes language predictions inspectable through geometric regions and temporal positions, providing an interface for analyzing how the model uses person dynamics (Appendix A.3.5).

## 5 Conclusion

We presented Conferencing Tokens, a compact representation of people in video conferencing that can be transmitted, reconstructed, and consumed directly by downstream models. Geometric initialization, rate–distortion learning, and causal entropy modeling establish a shared framewise decoding interface that connects to person-specific rendering and language models. Experiments show competitive reconstruction quality at low bitrates and substantially shorter language inputs while retaining meeting-state prediction accuracy close to the original RGB reference. These findings support a broader role for communication representations as reusable model interfaces that carry transmitted information into downstream analysis.

## AI USE STATEMENT

In this work, we used generative AI tools to assist with language polishing, including obtaining manuscript feedback through the LLM feedback tool provided by ICLR. We also used generative AI tools to assist with training-data annotation and answer refinement. We have reviewed all AI-assisted work and revised the manuscript accordingly. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

## REFERENCES

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, et al. Qwen3-VL technical report. arXiv:2511.21631, 2025. URL https://arxiv.org/abs/2511.21631.

Johannes Ballé, David Minnen, Saurabh Singh, Sung Jin Hwang, and Nick Johnston. Variational image compression with a scale hyperprior. In *International Conference on Learning Representations*, 2018. URL https://arxiv.org/abs/1802.01436.

Benjamin Bross, Ye-Kui Wang, Yan Ye, Shan Liu, Jianle Chen, Gary J. Sullivan, and Jens-Rainer Ohm. Overview of the versatile video coding (VVC) standard and its applications. *IEEE Transactions on Circuits and Systems for Video Technology*, 31(10):3736–3764, 2021. doi: 10.1109/TCSVT.2021.3101953.

Houwei Cao, David G. Cooper, Michael K. Keutmann, Ruben C. Gur, Ani Nenkova, and Ragini Verma. CREMA-D: Crowd-sourced emotional multimodal actors dataset. *IEEE Transactions on Affective Computing*, 5(4):377–390, 2014. doi: 10.1109/TAFFC.2014.2336244. URL https://pubmed.ncbi.nlm.nih.gov/25653738/.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, Jie Lei, Tengyu Ma, Baishan Guo, Arpit Kalla, Markus Marks, Joseph Greer, Meng Wang, Peize Sun, Roman Rädle, Triantafyllos Afouras, Effrosyni Mavroudi, Katherine Xu, Tsung-Han Wu, Yu Zhou, Liliane Momeni, Rishi Hazra, Shuangrui Ding, Sagar Vaze, Francois Porcher, Feng Li, Siyuan Li, Aishwarya Kamath, Ho Kei Cheng, Piotr Dollár, Nikhila Ravi, Kate Saenko, Pengchuan Zhang, and Christoph Feichtenhofer. SAM 3: Segment anything with concepts. arXiv:2511.16719, 2025. URL https://arxiv.org/abs/2511.16719.

Jean Carletta. Announcing the AMI meeting corpus. *The ELRA Newsletter*, 11(1):3–5, 2006. URL https://groups.inf.ed.ac.uk/ami/corpus/overview.shtml.

Bolin Chen, Zhao Wang, Binzhe Li, Rongqun Lin, Shiqi Wang, and Yan Ye. Beyond keypoint coding: Temporal evolution inference with compact feature representation for talking face video compression. In *Data Compression Conference (DCC)*, pp. 13–22, 2022. doi: 10.1109/DCC52660.2022.00009.

Hyomin Choi and Ivan V. Bajić. Scalable image coding for humans and machines. *IEEE Transactions on Image Processing*, 2022a. doi: 10.1109/TIP.2022.3160602. URL https://arxiv.org/abs/2107.08373.

Hyomin Choi and Ivan V. Bajić. Scalable video coding for humans and machines. arXiv:2208.02512, 2022b. URL https://arxiv.org/abs/2208.02512.

Yang Deng, Yukun Lai, and Paul L. Rosin. CCDb+: Enhanced annotations and multi-modal benchmark for natural dyadic conversations. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 5657–5666, 2025. doi: 10.1145/3746027.3755333.

Zhiyuan Fang, Jianfeng Wang, Xiaowei Hu, Lijuan Wang, Yezhou Yang, and Zicheng Liu. Compressing visual-linguistic model via knowledge distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021. URL https://openaccess.thecvf.com/content/ICCV2021/papers/Fang_Compressing_Visual-Linguistic_Model_via_Knowledge_Distillation_ICCV_2021_paper.pdf.

Ze Feng, Sen Yang, Boqiang Duan, Wankou Yang, and Jingdong Wang. EM-KD: Distilling efficient multimodal large language model with unbalanced vision tokens. arXiv:2511.21106, 2025. URL https://arxiv.org/abs/2511.21106. Accepted to AAAI 2026.

Xuan Gao, Chenglai Zhong, Jun Xiang, Yang Hong, Yudong Guo, and Juyong Zhang. Reconstructing personalized semantic facial NeRF models from monocular video. *ACM Transactions on Graphics*, 41(6), 2022. doi: 10.1145/3550454.3555501.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Yifan Hu, Rui Liu, Yi Ren, Xiang Yin, and Haizhou Li. UniTalker: Conversational speech-visual synthesis. In *Proceedings of the 33rd ACM International Conference on Multimedia*, 2025. doi: 10.1145/3746027.3755502. URL https://arxiv.org/abs/2508.04585.

Biao Jiang, Xin Chen, Wen Liu, Jingyi Yu, Gang Yu, and Tao Chen. MotionGPT: Human motion as a foreign language. In *Advances in Neural Information Processing Systems*, volume 36, 2023a. URL https://papers.nips.cc/paper_files/paper/2023/file/3fbf0c1ea0716c03dea93bb6be78dd6f-Paper-Conference.pdf.

Wei Jiang, Hyomin Choi, and Fabien Racapé. Adaptive human-centric video compression for humans and machines. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 1121–1129, 2023b. URL https://openaccess.thecvf.com/content/CVPR2023W/NTIRE/html/Jiang_Adaptive_Human-Centric_Video_Compression_for_Humans_and_Machines_CVPRW_2023_paper.html.

Tobias Kirschstein, Shenhan Qian, Simon Giebenhain, Tim Walter, and Matthias Nießner. NeRSemble: Multi-view radiance field reconstruction of human heads. *ACM Transactions on Graphics*, 42(4), 2023. doi: 10.1145/3592455.

Jiyoung Lee, Seungryong Kim, Sunok Kim, Jungin Park, and Kwanghoon Sohn. Context-aware emotion recognition networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10143–10152, 2019. URL https://openaccess.thecvf.com/content_ICCV_2019/html/Lee_Context-Aware_Emotion_Recognition_Networks_ICCV_2019_paper.html.

Jiahao Li, Wenxuan Xie, Zhaoyang Jia, Bin Li, Zongyu Guo, Xiaoyi Zhang, and Yan Lu. Ultra-fast neural video compression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026. URL https://arxiv.org/abs/2606.04410.

Jianglong Li, Jun Xu, Bingcong Lu, Zhengxue Cheng, Hongwei Hu, Ronghua Wu, and Li Song. Lightweight high-fidelity low-bitrate talking face compression for 3d video conference. In *2025 International Conference on Visual Communications and Image Processing (VCIP)*, pp. 1–5, 2025. doi: 10.1109/VCIP67698.2025.11396844. URL https://arxiv.org/abs/2601.21269.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 19730–19742, 2023. URL https://proceedings.mlr.press/v202/li23q.html.

Tianye Li, Timo Bolkart, Michael J. Black, Hao Li, and Javier Romero. Learning a model of facial shape and expression from 4D scans. *ACM Transactions on Graphics*, 36(6), 2017. doi: 10.1145/3130800.3130813.

Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Conference on Computer Vision and Pattern Recognition*, pp. 4104–4113, 2016. URL https://colmap.github.io/bibliography.html.

Johannes Lutz Schönberger, Enliang Zheng, Marc Pollefeys, and Jan-Michael Frahm. Pixelwise view selection for unstructured multi-view stereo. In *European Conference on Computer Vision*, 2016. URL https://colmap.github.io/bibliography.html.

Liam Schoneveld, Zhe Chen, Davide Davoli, Jiapeng Tang, Saimon Terazawa, Ko Nishino, and Matthias Nießner. SHeaP: Self-supervised head geometry predictor learned via 2d gaussians. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025. URL https://arxiv.org/abs/2504.12292.

Leqi Shen, Guoqiang Gong, Tao He, Yifeng Zhang, Pengzhang Liu, Sicheng Zhao, and Guiguang Ding. FastVID: Dynamic density pruning for fast video large language models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=2xS4VtpApy.

Luchuan Song, Pinxin Liu, Haiyang Liu, Zhenchao Jin, Yolo Yunlong Tang, Zichong Xu, Susan Liang, Jing Bi, Jason J. Corso, and Chenliang Xu. Bridging facial understanding and animation via language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17557–17567, June 2026. URL https://openaccess.thecvf.com/content/CVPR2026/papers/Song_Bridging_Facial_Understanding_and_Animation_via_Language_Models_CVPR_2026_paper.pdf.

Ting-Chun Wang, Arun Mallya, and Ming-Yu Liu. One-shot free-view neural talking-head synthesis for video conferencing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021. URL https://research.nvidia.com/publication/2021-06_one-shot-free-view-neural-talking-head-synthesis-video-conferencing.

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. doi: 10.1109/TIP.2003.819861. URL https://www.ece.uwaterloo.ca/~z70wang/publications/ssim.html.

Ziyi Wang, Yuhang Wu, Dongxu Piao, Xingyu Liu, Tianhui Zhou, and Miao Liu. MeetingToM: Evaluating multimodal llms on theory-of-mind reasoning in multi-party meetings. arXiv:2607.19235, 2026. URL https://arxiv.org/abs/2607.19235.

Adam Wieckowski, Jens Brandenburg, Tobias Hinz, Christian Bartnik, Valeri George, Gabriel Hege, Christian Helmrich, Anastasia Henkel, Christian Lehmann, Christian Stoffers, Ivan Zupancic, Benjamin Bross, and Detlev Marpe. VVenC: An open and optimized VVC encoder implementation. In *IEEE International Conference on Multimedia and Expo Workshops (ICMEW)*, pp. 1–2, 2021. doi: 10.1109/ICMEW53276.2021.9455944.

Jun Xiang, Xuan Gao, Yudong Guo, and Juyong Zhang. FlashAvatar: High-fidelity head avatar with efficient gaussian embedding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. URL https://ustc3dv.github.io/FlashAvatar/.

Liangbin Xie, Xintao Wang, Honglun Zhang, Chao Dong, and Ying Shan. VFHQ: A high-quality dataset and benchmark for video face super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 657–666, 2022. URL https://openaccess.thecvf.com/content/CVPR2022W/NTIRE/html/Xie_VFHQ_A_High-Quality_Dataset_and_Benchmark_for_Video_Face_Super-Resolution_CVPRW_2022_paper.html.

Shilin Xu, Xiangtai Li, Haobo Yuan, Lu Qi, Yunhai Tong, and Ming-Hsuan Yang. LLAVADI: What matters for multimodal large language models distillation. arXiv:2407.19409, 2024. URL https://arxiv.org/abs/2407.19409.

Xubing Ye, Yukang Gan, Xiaoke Huang, Yixiao Ge, Ying Shan, and Yansong Tang. VoCo-LLaMA: Towards vision compression with large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025. URL https://arxiv.org/abs/2406.12275.

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*, 2018. URL https://richzhang.github.io/PerceptualSimilarity/.

Shaolei Zhang, Qingkai Fang, Zhe Yang, and Yang Feng. LLaVA-Mini: Efficient image and video large multimodal models with one vision token. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/8400a2ec9ba85bfdb41eb2a584c0876d-Paper-Conference.pdf.

Jiaxing Zhao, Boyuan Sun, Xiang Chen, and Xihan Wei. Facial dynamics in video: Instruction tuning for improved facial expression perception and contextual awareness. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(16):13163–13171, 2026. doi: 10.1609/aaai.v40i16.38317. URL https://ojs.aaai.org/index.php/AAAI/article/view/38317.

# A Appendix

## A.1 Data and Implementation Details

### A.1.1 Data Construction and Splits

The candidate pool in Table 2 contains 11,000 clips from five sources. Tokenizer admission retains 10,925 clips, yielding 76,757 training and 6,297 validation windows. Sampling follows natural window frequencies. The ten geometry-evaluation identities are excluded from tokenizer training. Grouping follows the source-specific rules in Table 2.

Table 2: Candidate train/validation counts, accepted totals, and source-specific grouping.

| Source | Candidate train | Candidate val. | Accepted total | Grouping unit |
|---|---|---|---|---|
| AMI (Carletta, 2006) | 2,000 | 200 | 2,199 | Linked meetings/participants |
| CCDb+ (Deng et al., 2025) | 2,000 | 79 | 2,079 | Sessions/people |
| NeRSemble (Kirschstein et al., 2023) | 2,000 | 200 | 2,200 | Person |
| Open3DFaceVid (Song et al., 2026) | 1,000 | 100 | 1,099 | Synthetic prompt family |
| VFHQ (Xie et al., 2022) | 3,000 | 421 | 3,348 | Source video |
| Total | 10,000 | 1,000 | 10,925 | |

**Language data.** After filtering truncated responses and unusable inputs, the language data contain 13,426 question–answer pairs from 10,764 unique clips: 12,015 pairs from 9,795 clips for training and 1,411 pairs from 969 clips for validation, with no source-clip overlap. All five main SFT runs share 960 AI-generated replacement answers (877 training, 83 validation) and 12,466 original dense-RGB Qwen3-VL-8B teacher answers (Bai et al., 2025).

Replacement annotation uses deterministic sampling stratified by data source and original split. AI agents inspect the question and RGB frames with frame indices and timestamps, including uniformly sampled frames and frames near motion changes. Only answers are replaced; questions, videos, and splits remain unchanged. Annotation inputs exclude test answers and teacher/student predictions. Replacement answers are AI-generated and have not undergone comprehensive human validation.

**Preprocessing and isolation.** High-frame-rate videos are downsampled; training and motion losses use the actual frame intervals. Selected AMI training data have no meeting-group overlap with MeetingToM (Wang et al., 2026), and training sources exclude CAER, CREMA-D, and FEC-Bench. Isolation follows data sources and meeting groups; comprehensive cross-source near-duplicate and pretrained-exposure checks have not been performed.

### A.1.2 TOKENIZER ARCHITECTURE AND TRAINING

**Tokenizer training.** The final face-crop tokenizer family starts from the original converted SHeaP, an exact SVD factorization of its original local heads, and freshly initialized codec/entropy modules. Trainable modules are ViT blocks 10–11 (zero-based), local-output layer normalization, SVD analysis/synthesis projections and residuals, and entropy models. Original pose heads are frozen, with gradients propagated to the backbone. The visual encoder operates frame by frame, with temporal information used by the entropy models. Training uses 15 epochs, final-stage batch size 46, FP32, AdamW with frontend/codec learning rates $2\times10^{-7}/10^{-4}$, weight decay $10^{-4}$, gradient clipping 1, and seed 20260919. Final-stage $\lambda$ values are 12.5, 814, 7687, 24651, and 70000 in target-rate order. At epoch 10, the 3k run changes $\lambda$ from 50 to 12.5 and the 9k run from 253866 to 70000, continuing to epoch 15; the other three values remain unchanged. Supervision spans 16 frames; prefix context receives no gradient, invalid gaps break sequences, and tails use the last 16 frames.

**Quantization and coding.** The learned latent uses $k_t = \mathrm{round}(e_t/0.125)$ and $\hat{e}_t = 0.125k_t$, with 16 unit-step hyperlatents. Synthesis reconstructs 100 expression, six eye-rotation, and two eyelid coordinates from this latent. Separate MLPs predict hyperprior mean/scale, with history-Transformer corrections. Physical root/neck/jaw rotations use fixed-reference SO(3) grids, while translation uses a coordinate grid. Initial rotation/translation steps are 0.0005/0.00005; four learned group multipliers lie in $[1/4, 4]$. The entropy models use reconstructed history to condition symbol distributions. The hyperlatent factorized-logistic prior has fixed center zero and scale one.

Table 3: Tokenizer architecture. E/P denote latent and physical histories. All history inputs are reconstructed quantities.

| Module | Specification |
|---|---|
| Analysis / synthesis | Linear $768 \to 108$ (no bias) / $108 \to 108$ (original bias), each with a residual branch $x+\mathrm{SwiGLU}(\mathrm{LN}(x))$ of width $108 \to 216 \to 108$, whose output projection starts at zero. |
| SVD initialization | Stack original local heads as $W \in \mathbb{R}^{108\times768}$; retain full rank with $A = \sqrt{S}V^\top$, $B = U\sqrt{S}$, so $BA = W$. Residual output projections start at zero. |
| Hyperanalysis | LN and SwiGLU $108 \to 216 \to 16$; round with zero offset. |
| Hyper mean / scale | Separate $16 \to 128 \to 108$ SiLU MLPs; zero-initialized output layers. |
| E/P history | Last 16 frames, LN/SwiGLU input mapping to 128; two pre-LN Transformer layers, four heads, FFN 512/GELU, zero dropout; learned query/positions and final LN. |
| Distribution heads | Separate $128 \to 128 \to 108$ (E) or $128 \to 128 \to 12$ (P) SiLU heads for mean/log-scale; zero-initialized output layers. No P hyperprior. |

**Streaming encode/decode protocol.** History buffers retain up to 16 decoded frames; reset them and their validity masks at sequence start and every 64 frames. For each valid frame, encode/decode hyperlatents first, derive the latent-symbol distribution from them and the decoded E history, and then encode/decode $k_t$. Next derive the physical distribution from decoded P history and encode/decode the root, translation, neck, and jaw grid indices in that order, then finalize the shared per-frame range coder. Reconstruct $\hat{e}_t = 0.125k_t$ and physical controls using inverse SO(3) reference coordinates, then append only reconstructed histories. The history query uses zero current-pose controls. Both sides use the same rounded integer centers and deterministic scale-table quantization. Language adaptation takes $[\hat{e}_t, \hat{p}_t]$ at this point; reconstruction additionally applies synthesis to $\hat{e}_t$ and merges the resulting local parameters with $\hat{p}_t$ before FLAME. No future observations enter either entropy model.

**Dynamic payload accounting.** The measured bitrate is $R_{\mathrm{payload}} = 8B_{\mathrm{coded}}/\tau$, where $B_{\mathrm{coded}}$ counts bytes for both branches and hyperlatents, including escapes, coder finalization, and padding. Experimental file containers, network transport overhead, and static receiver assets are excluded.

**Tokenizer distortion.** $D = D_{\mathrm{param}} + 1.40283717(D_{\mathrm{local}} + D_{\mathrm{world}} + D_{\mathrm{velocity}})$. The direct non-expression groups in $D_{\mathrm{param}}$ have unit weights. Rotation losses are squared SO(3) angular errors normalized by squared RMS standard deviation; translation and eyelids use per-dimension standard-deviation-normalized MSE. Local/world vertex differences are divided by 0.01 m, squared, summed

over XYZ, and averaged. Local velocity uses actual frame intervals and a 0.1 m/s normalization. The reference geometry and motion are cached SHeaP estimates.

#### A.1.3 Language Adaptation Configuration

The implementation uses $d_e = 108$, $d_p = 12$, $d_z = 16$, and $d_{\mathrm{LM}} = 4096$. Each rate uses its 15-epoch tokenizer and a fresh Qwen3-VL-8B-Instruct (Bai et al., 2025) adapter/LoRA. The linear $120 \to 4096$ adapter includes a learned missing-input embedding (approximately 0.50M parameters). LoRA covers attention query/key/value/output projections in blocks 0–34, with rank 16, $\alpha = 32$, zero dropout, and approximately 14.91M parameters; block 35 remains frozen. AdamW uses adapter/LoRA learning rates $10^{-4}/(2\times10^{-5})$, weight decay 0.01, and clipping 1. Base weights use BF16; trainable weights, losses, and optimizer state use FP32. Physical/effective batch sizes are 2/16, with 751 updates per epoch for three epochs, seed 20260922, and constant learning rates. Answer-only cross-entropy includes EOS.

### A.2 Reconstruction Evaluation

#### A.2.1 Renderer and Baseline Implementation

**Renderer training.** Each identity/operating-point renderer is trained from scratch with seed 0 for 150k steps, with evaluation every 5k steps. The renderer training objective sums RGB Huber loss ($\delta = 0.1$), 40 times mouth Huber loss, 0.05 times VGG head-image loss (after 15k steps), and $\mathrm{mean} \| \mathrm{ReLU}(\mathrm{scale} - 0.6)\|$. Deformation Adam uses learning rate $10^{-4}$ and $\beta = (0.9, 0.999)$. Gaussian Adam rates are 0.0025/0.000125 for DC/remaining spherical harmonics, 0.05 for opacity, 0.005 for scale, and 0.001 for rotation, with $\epsilon = 10^{-15}$.

The implementation maps `torso_pose` to FLAME root rotation; neck is separate. Deformation conditions on 100 expression, three jaw, six eye, and two eyelid coordinates, plus canonical positional embeddings, and predicts Gaussian position/rotation/scale residuals.

**Baselines.** VVenC uses commit `b1cfb2ad495a`, native $512 \times 512$, medium preset, GOP-8 low delay, and disabled reordering/lookahead/MCTF/BIM. QPs are 32, 37, 42, 47, 52, 57, plus 50, 60, and 63. DCVC-UF uses commit `cbdae87a5445`, official CVPR2026 image/low-delay video checkpoints, skip 0.15, QPs 0/1/2/4, and reset interval 64. FV2V uses the OSFW implementation (commit `9511d25a841a`, Vox256NewSPADE), Lanczos downsampling to $256 \times 256$ and bicubic upsampling to $512 \times 512$, VTM 22.2 initial-frame coding, and differential Exp-Golomb/adaptive arithmetic coding of 57 motion parameters. Li et al. '25* (adapted) codes the first $D \in \{10, 20, 30, 40, 50\}$ expression coordinates plus 20 other parameters with 10-bit quantization and DPCM/adaptive Rice coding, with a dedicated renderer per point.

#### A.2.2 Evaluation and Checkpoint Selection

**Data and metrics.** Reconstruction uses Obama, id1, id2, id4, id5, id6, id7, and id8 from NeRF-BlendShape (Gao et al., 2022), with $512 \times 512$ preprocessing, SAM 3 (Carion et al., 2025) foreground alpha, and SHeaP parameters. The final 500 frames per sequence form the test set; the remainder train its renderer (1,901 frames for Obama).

All methods share frame IDs, train/test splits, white-background compositing, and fixed ground-truth upper-body masks. GT is alpha-composited onto white and rounded to uint8; decoded RGB is clamped and rounded to uint8. PSNR uses full-range floating BT.709 YUV420 with 2-by-2 box chroma downsampling, all-four-pixel foreground support for chroma, and Y:U:V MSE weights 4:1:1; RGB SSIM (Wang et al., 2004) uses a 7-by-7-eroded foreground mask, and spatial LPIPS (Zhang et al., 2018) uses VGG with foreground masking. Results are macro-averaged across people. All Proposed and Li et al. '25* renderers, including temporal ablations, follow the common selection rule below.

**Renderer checkpoint selection.** For each identity, method, and operating point, checkpoint selection uses the last 50 frames of its original renderer training interval (Table 4). These frames belong to the training interval and are disjoint from the final 500 test frames. The selection criterion is maximum arithmetic-mean PSNR over these 50 frames among 30 checkpoints from 5k to 150k steps at

5k intervals, with ties resolved in favor of the earlier step. The same selected checkpoint supplies test PSNR, SSIM, LPIPS, and qualitative reconstructions. E+P at 6 kbps uses the matching main renderer and selection. Tables 5 and 9 report the selected steps.

Table 4: Inclusive source-frame IDs for training-subset selection and final testing. Every selection interval contains 50 previously trained-on frames; every test interval contains 500 frames.

| Identity | Training selection | Test |
|---|---|---|
| Obama | 1851–1900 | 1901–2400 |
| ID 1 | 2704–2753 | 2754–3253 |
| ID 2 | 2891–2940 | 2941–3440 |
| ID 4 | 3094–3143 | 3144–3643 |
| ID 5 | 2450–2499 | 2500–2999 |
| ID 6 | 2450–2499 | 2500–2999 |
| ID 7 | 3496–3545 | 3546–4045 |
| ID 8 | 3047–3096 | 3097–3596 |

Table 5: Selected renderer steps in thousands. Proposed columns are target kbps; Li columns are expression dimensions. All entries use the same training-subset PSNR rule.

| | Proposed | | | | | Li et al. '25* | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Identity | 3 | 4.5 | 6 | 7.5 | 9 | 10 | 20 | 30 | 40 | 50 |
| Obama | 90 | 145 | 145 | 125 | 140 | 145 | 145 | 140 | 145 | 145 |
| ID 1 | 145 | 150 | 140 | 140 | 130 | 145 | 135 | 85 | 125 | 135 |
| ID 2 | 95 | 140 | 135 | 145 | 145 | 145 | 150 | 145 | 130 | 120 |
| ID 4 | 100 | 125 | 140 | 120 | 145 | 65 | 115 | 130 | 135 | 135 |
| ID 5 | 135 | 130 | 150 | 125 | 135 | 130 | 135 | 135 | 120 | 115 |
| ID 6 | 125 | 115 | 145 | 125 | 110 | 130 | 125 | 125 | 150 | 150 |
| ID 7 | 70 | 110 | 140 | 40 | 50 | 50 | 55 | 55 | 125 | 150 |
| ID 8 | 110 | 150 | 150 | 130 | 125 | 150 | 140 | 140 | 150 | 150 |

### A.2.3 Per-Subject Rate–Distortion Results

We compare Proposed with the adapted Li et al. '25* baseline using all five measured operating points per method and subject. For each subject, piecewise cubic Hermite interpolation (PCHIP) models PSNR as a function of natural-log bitrate. We integrate the difference (Proposed minus baseline) over that subject's common log-rate interval and divide by its width, without extrapolation. The reported $+0.5344$ dB is the equal mean of the eight subject-level integrals in Table 6.

Table 6: Per-subject BD-PSNR relative to Li et al. '25*, with the common measured bitrate interval used for each integral.

| Subject | Common interval (kbps) | BD-PSNR (dB) |
|---|---|---|
| Obama | 5.6400–8.6732 | $+0.2946$ |
| ID 1 | 5.0544–9.1516 | $+1.1442$ |
| ID 2 | 4.9204–9.0456 | $+0.5931$ |
| ID 4 | 5.2308–8.7028 | $+1.2061$ |
| ID 5 | 5.4088–8.6960 | $-0.0199$ |
| ID 6 | 5.5128–8.4780 | $+0.3116$ |
| ID 7 | 5.5520–9.7364 | $+0.8547$ |
| ID 8 | 5.0384–8.7500 | $-0.1090$ |
| Equal-weight mean | — | $+0.5344$ |

### A.2.4 Static Receiver Assets

Dynamic payload rates exclude static receiver assets and network/file-container overhead. Table 7 reports FP32 tensor content for the selected 6 kbps-target renderers: all eight identities have 14,876

Gaussians and the same payload size. The count excludes training state, file packaging, and shared runtime resources.

Table 7: Static receiver asset content per identity for the selected Proposed renderers.

| Component | FP32 tensor bytes |
|---|---|
| Gaussian attributes | 3,332,224 |
| Deformation network | 1,493,032 |
| Static shape | 1,200 |
| Total | 4,826,456 |

The total is 4.826 MB per identity in decimal units. As a separate historical reference, archived VCIP assets for four identities (ID 2/4/5/8), each with five D10–D50 configurations, average 0.721 MB after attribute compression, FP16 deformation-network storage, and gzip. This reference excludes shape and shared resources and is a compressed asset measurement from another implementation, not from ours.

#### A.2.5 Temporal Entropy-Context Comparison Protocol

**Configurations.** The four history configurations in Section 4.3 share the 15-epoch training configuration, seed 20260919, final batch size 46, face-crop data, distortion loss, and SHeaP/SVD initialization described in Appendix A.1.2. History masking retains network capacity and the current-frame hyperprior.

**Independent geometry.** The ten identities are excluded from tokenizer training: development 030, 226, 227, 124, 097; final test 175, 085, 018, 038, 240. Final-test geometry does not select tokenizer weights. COLMAP (Schönberger & Frahm, 2016; Schönberger et al., 2016) reconstructs reference geometry independently from NeRSemble multiview images. Evaluation covers 58 frames, excluding the same two occluded frames of identity 097 for every model. Each model/frame receives its own seven-DoF similarity fit (scale, rotation, translation), initialized from its landmarks and refined by similarity ICP using a fixed-seed sample of 512 points. All eligible COLMAP points are scored by nearest-face-triangle distance in mm, averaged over points, frames, then identities. Intrinsic local meshes remove root, neck, and translation effects, so this tests shape rather than pose. Teacher-derived transforms are not used for this metric. Table 8 retains separate development/test means. Rates count combined payload and escape bytes at a 25-fps equivalent, excluding shape and headers; geometry and renderer rates refer to different sequences.

**Reconstruction.** ID 1 and ID 2 use 2,754 and 2,941 training frames, respectively, with the final 500 frames per identity used for evaluation. Dedicated renderers follow Appendix A.2.1 and the checkpoint rule in Appendix A.2.2; the 6 kbps-target E+P configuration uses its matching main renderer. Table 9 reports the per-identity results.

### A.3 Understanding Evaluation

#### A.3.1 Inputs, Sampling, and Scoring

For the STATE comparison in Section 4.1, eleven examples are excluded because of missing usable faces, boundary crops, back-facing subjects, or multiple-person ambiguity. RGB uses the full closeup, while tokens use a fixed center/face crop. Inputs contain the target-agent designation and task text without audio. Accuracy and unweighted class-macro F1 are percentages. Greedy decoding uses at most 96 new tokens and the official JSON output format, without chain-of-thought repair.

**Sampling and cost.** Sparse RGB uses ten frames at 2 fps; dense RGB uses 125 frames at 25 fps, duplicating the final frame to reach 126. RGB is resized from $720 \times 576$ to $384 \times 307$ and black-padded to $384 \times 384$. Timestamp/control structure is retained, and shared-text three-axis positions align with dense RGB. Modality input-position counts exclude timestamps, control, text, and generated outputs. Rates in Table 1 are computed over the common 2,945-second duration. Sparse/dense

Table 8: History-ablation geometry by split: distance $D$ (mm) and measured 25-fps-equivalent payload rate $R$ (kbps). Dev/Test average five identities each; All averages ten.

| | No history | | E history | | P history | | E+P history | |
|---|---|---|---|---|---|---|---|---|
| Split | $D$ | $R$ | $D$ | $R$ | $D$ | $R$ | $D$ | $R$ |
| Dev | 1.7313 | 6.760 | 1.7319 | 6.233 | 1.7322 | 6.215 | 1.7336 | 5.618 |
| Test | 1.9655 | 6.659 | 1.9684 | 6.089 | 1.9681 | 6.173 | 1.9632 | 5.568 |
| All | 1.8484 | 6.710 | 1.8501 | 6.161 | 1.8501 | 6.194 | 1.8484 | 5.593 |

Table 9: Per-identity history-ablation results after training-subset PSNR selection. Steps are in thousands. These rows supply the two-identity averages in Figure 5. Rates are kbps.

| ID | History | Step | Rate | PSNR | SSIM | LPIPS |
|---|---|---|---|---|---|---|
| 1 | No history | 145 | 6.1772 | 28.7532 | 0.8306 | 0.1962 |
| 1 | E history | 140 | 5.6924 | 29.1029 | 0.8381 | 0.1889 |
| 1 | P history | 140 | 5.7796 | 28.8323 | 0.8335 | 0.1930 |
| 1 | E+P history | 140 | 5.2920 | 28.8857 | 0.8360 | 0.1912 |
| 2 | No history | 115 | 6.2816 | 27.5758 | 0.7932 | 0.2845 |
| 2 | E history | 135 | 5.8716 | 27.6449 | 0.7952 | 0.2784 |
| 2 | P history | 140 | 5.8124 | 27.1628 | 0.7907 | 0.2824 |
| 2 | E+P history | 135 | 5.3068 | 27.3977 | 0.7935 | 0.2783 |

VVenC inputs share the complete elementary stream, including codec headers; Conferencing Tokens count entropy-coded payload as defined in Section 3.2.

### A.3.2 Input-Dependence Diagnostics

Table 10 uses the main models: the 15-epoch tokenizer and three-epoch SFT model at update 2,253 for each rate. Cross-session mismatch fixes the model weights, question, transcript, target-agent designation, and positional structure, replacing only the complete 125-frame person-token sequence and validity mask with those from another session. A hash of the clip ID determines the donor without outcome-based selection. At $\lambda = 7687$ (6.30 kbps), the paired accuracy decrease is 18.51 percentage points, with a 95% interval of [14.77, 22.18] from 5,000 meeting-group bootstrap resamples.

Table 10: Input dependence on STATE with fixed weights within each row. Each pair reports accuracy / macro-F1 (%). Mismatch replaces the person sequence across sessions; hard-off blocks attention access to visual keys/values.

| $\lambda$ | Matched | Mismatched | Hard-off |
|---|---|---|---|
| 12.5 | 58.57 / 22.35 | 44.48 / 17.28 | 12.39 / 6.80 |
| 814 | 54.84 / 20.52 | 42.28 / 16.44 | 9.68 / 4.64 |
| 7687 | 59.25 / 21.91 | 40.75 / 15.29 | 11.38 / 6.25 |
| 24651 | 57.89 / 21.00 | 40.92 / 14.46 | 17.49 / 8.33 |
| 70000 | 58.74 / 23.13 | 42.11 / 14.87 | 54.50 / 21.28 |

Hard-off blocks nonvisual queries from attending to visual keys/values in all 36 layers during prefill and generation, retaining the inputs and modality positions; this differs from text-only inference. Its effect varies across rates (Table 10). As a separate reference with different, unadapted model weights, Qwen3-VL-8B with the visual structure removed and text positions recomputed obtains 280/589 correct (47.54% accuracy, 22.51% macro-F1).

### A.3.3 Class-Level Results

Table 11 shows that Proposed performs similarly to the RGB references on focused listening and active engagement, with uneven performance on the other categories. The RGB references also

perform weakly on some minority categories. Matched-input gains are concentrated in focused listening and active engagement.

Table 11: STATE per-class F1 (%) on the same 589 examples. Support is the true-label count. Proposed uses $\lambda = 7687$; Mismatched changes only its person-token sequence across sessions. RGB references receive no additional task SFT. Zero-prediction classes have F1 zero.

| Class | Support | Sparse RGB F1 | Dense RGB F1 | Proposed F1 | Mismatched F1 |
|---|---|---|---|---|---|
| Disengaged withdrawal | 49 | 17.91 | 13.56 | 0.00 | 0.00 |
| Hesitation | 36 | 13.64 | 5.26 | 5.00 | 10.00 |
| Confused bewilderment | 27 | 0.00 | 0.00 | 0.00 | 0.00 |
| Cognitive conflict | 28 | 6.90 | 0.00 | 0.00 | 0.00 |
| Supportive endorsement | 56 | 6.67 | 0.00 | 6.45 | 0.00 |
| Focused listening | 168 | 63.86 | 63.24 | 64.67 | 43.74 |
| Active engagement | 225 | 74.36 | 74.61 | 77.23 | 53.28 |
| Macro-F1 | — | 26.19 | 22.38 | 21.91 | 15.29 |

Proposed prediction counts in row order are 0, 4, 0, 0, 6, 299, and 280 (589 total). Class supports also sum to 589; macro-F1 is the unweighted mean of the seven class F1 scores.

### A.3.4 Cross-Scenario Emotion Classification

Table 12 evaluates a fixed CREMA-D (Cao et al., 2014) subset of 337 examples and 146 input-QC-selected CAER (Lee et al., 2019) examples from FEC-Bench (Zhao et al., 2026), using the main experimental configuration (Section 4.1). A performance gap remains between Conferencing Tokens and the RGB references in this cross-scenario evaluation.

Table 12: Cross-scenario classification performance (%). RGB predictions are frozen and shared across rate comparisons. Proposed row labels are target rates, not measured rates.

| | CREMA-D (337) | | FEC-Bench CAER (146) | |
|---|---|---|---|---|
| Input (target kbps) | Acc. | Macro-F1 | Acc. | Macro-F1 |
| Sparse RGB | 36.50 | 33.64 | 48.63 | 30.35 |
| Dense RGB | 33.53 | 31.36 | 53.42 | 35.39 |
| Proposed (3) | 19.58 | 10.28 | 37.67 | 16.21 |
| Proposed (4.5) | 19.88 | 10.57 | 36.30 | 17.04 |
| Proposed (6) | 20.47 | 12.25 | 36.30 | 15.17 |
| Proposed (7.5) | 21.96 | 14.75 | 34.93 | 15.60 |
| Proposed (9) | 20.77 | 12.69 | 36.30 | 14.30 |

### A.3.5 Geometric Sensitivity Analysis Protocol

**Inputs and score.** Figure 6 illustrates local answer sensitivity for one person under selected jaw-opening trajectories. The 6 kbps-target SFT model receives 108 decoded latent coordinates and 12 physical coordinates without standardization. Each sequence contains 50 frames at 25 fps. Only the jaw $x$ axis-angle control varies through an initial hold, smooth opening, and final hold; other inputs, shape, decoder, camera, and renderer remain fixed. Opening amplitudes follow the sweep shown in the figure. The question is *From the beginning to the end of this clip, does the mouth become more open? Answer only Yes or No.* The score is the raw first-answer-token Yes-minus-No logit.

**Gradient visualization.** We project the input gradient onto latent/jaw coordinates and push it through the geometry decoder as in Equation 7. The visualization uses one shared gradient normalization and a shared displacement-magnitude scale set at the 99th percentile over the original full visualization collection. These fixed scales are applied to the selected displayed frames, with consistent arrow enlargement and anchor subsampling.

### A.3.6 JOINT QUALITATIVE COMPARISON

Figure 4 uses corresponding two-second clips. Sparse and dense RGB consume 4 and 50 VVenC-reconstructed frames, respectively, with 288 and 3,600 visual tokens; Proposed uses 50 conferencing tokens. Rates are measured over the corresponding person's full 500-frame sequence under Section 3.2. Renderer checkpoints follow Appendix A.2.2.

### A.3.7 LANGUAGE ADAPTATION OF CODED FACIAL PARAMETERS

We adapt parameters coded by Li et al. '25* (Li et al., 2025) for language input at D10 and D50. Inputs are the original SHeaP/FLAME parameters after the current camera-coordinate conversion and Li encoding/decoding, rather than parameters recovered by the Proposed geometry head. Both parameter-input runs use three epochs (2,253 updates), an effective batch size of 16, and the final checkpoint without intermediate validation selection. Evaluation uses the same 589 STATE examples and retains the existing 11 exclusions. Rates in Table 13 measure dynamic payload; each input uses 125 modality positions.

Table 13: Language adaptation of coded facial parameters and Conferencing Tokens on STATE ($N = 589$). All rates are measured dynamic payload rates. Adapted Li et al. '25* denotes our language adaptation of the coded parameter inputs.

| Representation / setting | Measured payload rate (kbps)↓ | STATE accuracy (%)↑ | Macro-F1 (%)↑ | Modality input positions↓ |
|---|---|---|---|---|
| Adapted Li et al. '25* (D10) | 4.515 | 55.86 | 20.05 | 125 |
| Adapted Li et al. '25* (D50) | 11.233 | 58.57 | 21.11 | 125 |
| Proposed ($\lambda = 12.5$) | 4.24 | 58.57 | 22.35 | 125 |
| Proposed ($\lambda = 814$) | 5.02 | 54.84 | 20.52 | 125 |
| Proposed ($\lambda = 7687$) | 6.30 | 59.25 | 21.91 | 125 |
| Proposed ($\lambda = 24651$) | 7.71 | 57.89 | 21.00 | 125 |
| Proposed ($\lambda = 70000$) | 10.41 | 58.74 | 23.13 | 125 |

The coded facial parameters and Conferencing Tokens yield similar STATE performance across the evaluated settings, suggesting that the learned compact representation retains comparable task-relevant information. With mismatched person inputs, accuracy decreases to 44.14% for D10 and 41.43% for D50, indicating that the coded parameter inputs also contribute to prediction.

## A.4 ALTERNATIVE ADAPTATION OBJECTIVES

We additionally examine response-level and intermediate-feature distillation as exploratory alternatives for language adaptation, motivated by prior multimodal distillation work (Fang et al., 2021; Xu et al., 2024; Feng et al., 2025). The objectives and controlled comparisons are detailed below.

### A.4.1 EXPLORATORY OBJECTIVE COMPARISON

Table 14 uses the earlier epoch-3, 6 kbps-target codec and original teacher answers, whereas the main SFT results use 15-epoch codecs and 960 agent-replaced answers. Within this comparison, all six runs share data/codec version, seed, effective batch size 16, three epochs, and learning rates; physical batch sizes and forward implementations differ.

In this comparison, additional KD or hidden supervision provides no consistent benefit, so the main experiments use SFT.

### A.4.2 OBJECTIVE DEFINITIONS

Each enabled loss has weight one. KD is teacher-to-student forward KL at temperature 2, scaled by $2^2$, using the top 512 teacher entries plus a tail bucket. Hidden uses block-34 outputs before final RMSNorm at valid prefill text positions. With $\Delta_s = h_s^{\text{normal}} - h_s^{\text{hardoff}}$ and $\Delta_t = h_t^{\text{normal}} - h_t^{\text{hardoff}}$,

$$\mathcal{L}_{\text{Hidden}} = \text{mean}\left[\left(\frac{\Delta_s - \Delta_t}{\max(\text{RMS}_t, 10^{-6})}\right)^2\right], \tag{8}$$

Table 14: Exploratory language-adaptation objectives on two fixed test subsets with the earlier face-crop 6 kbps-target codec and three adaptation epochs. Each entry reports correct answers / test size and accuracy. Hidden denotes layer-34 residual supervision under the normal-hardoff configuration; each row is a single training run.

| Objective | STATE: correct / 589 (%) | CREMA-D: correct / 337 (%) |
|---|---|---|
| SFT | 341 / 589 (57.89) | 70 / 337 (20.77) |
| SFT+KD | 335 / 589 (56.88) | 71 / 337 (21.07) |
| Hidden-only | 329 / 589 (55.86) | 69 / 337 (20.47) |
| SFT+Hidden | 341 / 589 (57.89) | 72 / 337 (21.36) |
| SFT+KD+Hidden | 337 / 589 (57.22) | 69 / 337 (20.47) |
| KD+Hidden | 334 / 589 (56.71) | 73 / 337 (21.66) |

where $\mathrm{RMS}_t$ is the elementwise RMS of the cached teacher residual. Hardoff preserves token positions while blocking nonvisual queries from visual keys/values. Both student branches receive gradients, the teacher is frozen, and no hidden-loss warmup is used.