

– *Supplementary Material* –

SYNERGIZING MOTION AND APPEARANCE: MULTI-SCALE COMPENSATORY CODEBOOKS FOR TALKING HEAD VIDEO GENERATION

Anonymous authors

Paper under double-blind review

1 SETTINGS.

We perform multi-scale compensation across $N = 4$ scales. We employ the keypoint-based motion flow estimator from Siarohin et al. (2019). The multi-scale motion flows are estimated at size 64×64 . We use convolution layers to encode the motion flows into a latent motion flow space of size $32 \times 32 \times 32$ and set the multi-scale motion codebook size as $K = 1024$ and $d_m = 32$. We also use convolution layers to decode the quantized motion flow features, while adopt the motion flow updater in Tao et al. (2024) as our motion flow residual decoder. We employ the image encoder and decoder architecture from Esser et al. (2021) and further encode the multi-scale appearance features into size $32 \times 32 \times 256$. The multi-scale appearance codebook size is set to $T = 1024$ and $d_a = 256$. For the training objective, we use the perceptual loss in Siarohin et al. (2019) and the L1 loss as the image reconstruction loss, and we set the loss weights as $\lambda_{adv} = 0.8$, $\lambda^1 = 0.5$, $\lambda_{recon,m} = 32$ and $\beta = 0.25$. We train the whole framework end-to-end with a learning rate of 8×10^{-5} and a batch size of 16 for 250K iterations on four NVIDIA RTX 3090 GPUs.

2 ADDITIONAL EXPERIMENTAL RESULTS

Table 1: Ablation study on the codebook allocation scheme. We present the results of different codebook splitting settings.

| | FID ↓ | PSNR ↑ | \mathcal{L}_1 ↓ | LPIPS ↓ | AKD ↓ | AED ↓ |
|-----------------------------|--------------|--------------|-------------------|---------------|---------------|---------------|
| Sharing all codes | 43.23 | 25.12 | 0.0359 | 0.1860 | 1.2124 | 0.1065 |
| Splitting the codes equally | 42.52 | 25.20 | 0.0358 | 0.1857 | 1.1893 | 0.1075 |
| Code Allocation (Ours) | 43.15 | 25.30 | 0.0355 | 0.1846 | 1.2039 | 0.1071 |

For both the motion and appearance codebooks, we propose a novel code allocation scheme that assigns different codes to corresponding scales. This allows certain codes to be shared across multiple scales, facilitating the transfer of information between them. To assess the effectiveness of our codebook structure, we conduct an ablation study, as shown in Tab. 1. We compare two alternative strategies: sharing all codes across all scales, and splitting the codes equally among the scales. As demonstrated in Tab. 1, our code allocation scheme achieves the best overall performance considering all the metrics. These results highlight the effectiveness and superiority of our proposed codebook allocation scheme.

REFERENCES

- Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *CVPR*, 2021.
- Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. First order motion model for image animation. *NeurIPS*, 2019.

054 Jiale Tao, Shuhang Gu, Wen Li, and Lixin Duan. Learning motion refinement for unsupervised face
055 animation. *NeurIPS*, 2024.
056
057
058
059
060
061
062
063
064
065
066
067
068
069
070
071
072
073
074
075
076
077
078
079
080
081
082
083
084
085
086
087
088
089
090
091
092
093
094
095
096
097
098
099
100
101
102
103
104
105
106
107