Robust Video Portrait Reenactment via Personalized Representation Quantization

Kaisiyuan Wang[1], Changcheng Liang[2], Hang Zhou[3*], Jiaxiang Tang[4], Qianyi Wu[5], Dongliang He[3], Zhibin Hong[3], Jingtuo Liu[3], Errui Ding[3], Ziwei Liu[6], Jingdong Wang[3]

[1]The University of Sydney
[2]Xidian University
[3]Baidu Inc.
[4]Peking University
[5]Monash University
[6]S-Lab, Nanyang Technological University

kaisiyuan.wang@sydney.edu.au, ccliang_xd@163.com, tjx@pku.edu.cn, qianyi.wu@monash.edu,
{zhouhang09, hedongliang01, liujingtuo, dingerrui, wangjingdong}@baidu.com,
Zhib.hong@gmail.com, zwliu.hust@gmail.com

Abstract

While progress has been made in the field of portrait reenactment, the problem of how to produce high-fidelity and robust videos remains. Recent studies normally find it challenging to handle rarely seen target poses due to the limitation of source data. This paper proposes the Video Portrait via Non-local Quantization Modeling (VPNQ) framework, which produces pose- and disturbance-robust reenactable video portraits. Our key insight is to learn position-invariant quantized local patch representations, then build a mapping between simple driving signals and local textures with non-local spatial-temporal modeling. Specifically, instead of learning a universal quantized codebook, we identify that a personalized one can be trained to preserve desired position-invariant local details. Then, a simple representation of projected landmarks can be used as sufficient driving signals to avoid 3D rendering. In the following, we employ a carefully designed Spatio-Temporal Transformer to predict reasonable and temporally consistent quantized tokens from the driving signal. The predicted codes can be decoded back to robust and high-quality videos. Comprehensive experiments have been conducted to validate the effectiveness of our approach.

Introduction

Synthesizing and driving video portraits aims to animate the target portrait with similar movements to the driving video, which enables applications in video editing, filmmaking, visual dubbing and digital human creation. Though several studies pursue vivid talking heads from only one or a few samples (Zakharov et al. 2019; Siarohin et al. 2019; Zakharov et al. 2020; Wang, Mallya, and Liu 2021; Zhou et al. 2019; Wang et al. 2022; Ji et al. 2022; Doukas, Zafeiriou, and Sharmanska 2020; Zhou et al. 2021), they tend to create visible distortions and identity drift, which makes their results less realistic and not applicable in most real-world applications. On the other hand, researchers seek person-specific modeling (Suwajanakorn, Seitz, and Kemelmacher-Shlizerman 2017; Kim et al. 2018; Ji et al. 2021; Kim et al. 2019; Gafni et al. 2021; Guo et al. 2021) for driving photo-realistic video portraits. The problem is mostly formulated as synthesizing textures according to the *geometry or semantic guidance* provided by the input driving signals.

Previous works can be divided into two categories: **1)** GAN-based methods where the geometry guidance is either rendered 3D faces or landmarks. A certain studies (Kim et al. 2018, 2019; Lahiri et al. 2021; Thies et al. 2020; Li et al. 2021) leverage rendered 3D faces (Blanz and Vetter 1999) and Generative Adversarial Networks. Their facial parts are rendered to be realistic, but the torso and interior teeth depend completely on generative models. Differently, Live Speech Portraits (LSP) (Lu, Chai, and Cao 2021) verifies that landmarks projected on 2D space can be directly mapped to talking portraits with GANs, but their results are unstable. **2)** Volume render-based methods. Recently, NeRF (Mildenhall et al. 2020) and other volume rendering approaches (Gafni et al. 2021; Guo et al. 2021; Zheng et al. 2022; Grassal et al. 2022) have shown success in producing portrait avatars. However, these methods require tedious preprocessing and are extremely time-consuming when building a single subject.

Moreover, the robustness of previous methods on in-the-wild monocular videos, which are the most common data sources, is not thoroughly considered. GAN-based methods build a direct mapping between the driving signals and targets. Their synthesized texture is strongly coupled with geometry guidance, which inevitably results in undesirable artifacts when processing unseen driving signals (*e.g.*, novel poses, different scales, and distorted signals). Meanwhile, NeRF-based methods are also sensitive to input driving signals, where a slight subject movement or inaccurate camera pose estimation may lead to the degradation of the generation quality. Therefore, how to synthesize robust and photo-realistic portrait videos remains a challenge.

To cope with these issues, we present a novel framework named **Video Portrait via Non-local Quantization Modeling (VPNQ)**, which achieves robust and high-fidelity video portrait reenactment even under challenging scenarios. The task is to drive a target portrait with an arbitrary portrait video. Our key insight is to *learn position-*

*Corresponding author.

invariant quantized representations and build a mapping between simple driving signals and local textures with non-local spatial-temporal modeling. Our method is inspired by the recent success of the vector-quantized generative model (Van Den Oord, Vinyals et al. 2017; Esser, Rombach, and Ommer 2021). We identify two intrinsic properties that make quantized representations suitable for synthesizing video portraits: **1)** The learned codebook can be regarded as a position-invariant texture dictionary, enabling the disentanglement between specific geometry locations and textures. **2)** The quantized generative model always encodes the input signal into a set of discrete quantized codes from the codebook, even if the input signal is slightly modified or has rarely been seen.

Specifically, different from previous methods that learn a universal quantized representation on a large scale of data (Esser, Rombach, and Ommer 2021; Chang et al. 2022; Gu et al. 2022), we find that such learned representation cannot recover person-specific details. Thus, we propose to learn a personalized quantized representation with an autoencoder from the target portrait only so that the learned codebook stores position-invariant and personalized delicate textural details. Given the locality of the textures stored in each patch, they will be less sensitive to the slight scale changes of the input information. As a result, we adopt the simplest setting as LSP to use the projected 2D landmarks as geometry guidance, which avoids the time-consuming neural rendering or volume rendering procedure.

The next question is how to map the driving signals to the decoupled personalized quantized textures. Given the discrete codebook learned above, the generative task can be reformulated into a classification (*i.e.*, code prediction) problem on spatial locations. However, CNNs that perform locally are not suitable for building the non-local correlation between quantized patches. Moreover, the temporal inconsistency problem has been witnessed (Hong et al. 2022; Zhou et al. 2022) in VQGAN for video generation.

To address the issues discussed above, we take advantage of Vision Transformers (ViT) (Dosovitskiy et al. 2021; Bao, Dong, and Wei 2021) and propose a *Spatio-Temporal Code Transformer* to operate on consecutive frames. Concretely, the Transformer directly processes all encoded features from a short video clip and predicts the corresponding codes. As the self-attention mechanism is naturally suitable for non-local spatial and temporal information modeling, the Transformer complements the reorganization of quantized local patches for high-fidelity video portraits.

Our contributions are summarized below: **1)** We propose **Video Portrait via Non-local Quantization Modeling (VPNQ)**, a novel framework that synthesizes robust and high-quality video portraits even from unseen and corrupted driving signals. **2)** We propose using personalized representation quantization by modeling local textural patches for a specific target. Such practice decouples driving signals and specific textures in portrait reenactment. **3)** We propose *Spatio-Temporal Code Transformer* by extending the previous image modeling into video modeling to predict reliable and temporally consistent codes for high-fidelity videos.

Related Works

Face Reenactment. By modeling the motion as latent representations, recent studies (Siarohin et al. 2019; Zakharov et al. 2019; Wang, Mallya, and Liu 2021; Wang et al. 2022) tend to generate video portraits based on only one or few source frames in a warping-based paradigm. However, these methods usually suffer from identity distortion and low generation quality.

In order to achieve stable and photo-realistic video portraits, the early works (Kim et al. 2018, 2019) focus on developing personalized models which rely on the 3DMM face model (Blanz and Vetter 1999) for human head rendering and a 2D generative model for the torso and background synthesis. While LSP (Lu, Chai, and Cao 2021) proposes to use 2D facial landmarks projected from 3D geometry as the driving guidance. However, the generated results are quite sensitive to the input landmarks. The latest studies (Gafni et al. 2021; Guo et al. 2021; Zheng et al. 2022; Grassal et al. 2022) take volume rendering into consideration. Nerface (Gafni et al. 2021) builds a talking head system by combining a dynamic radiance field with a low-dimensional morphable model. Similarly, NHA (Grassal et al. 2022) also presents a hybrid representation, which includes a morphable model and two feed-forward networks for vertex offset and expression texture prediction.

Differently, our VPNQ avoids the time-consuming volume rendering as well as tedious preprocessing and develops a robust face reenactment framework by using the simplest 2d landmarks driving strategy.

Quantization-based Image Modeling. Due to the success of natural language processing, increasing attention has been paid to image modeling based on Transformer networks, among which quantized image modeling (*e.g.*, VQVAE (Van Den Oord, Vinyals et al. 2017) and VQGAN (Esser, Rombach, and Ommer 2021)) has shown great potential in synthesizing high-resolution images with rich content.

Typically, these quantized generative models include an autoencoder architecture and a learnable codebook. The core of these approaches is to replace the discrete representations from the encoder with the code from the learned codebook (*a.k.a*, the quantization process) before decoding. The existing approaches (Esser, Rombach, and Ommer 2021; Chang et al. 2022; Gu et al. 2022; Esser et al. 2021) usually learn a universal codebook via training on large-scale datasets. VQGAN (Esser, Rombach, and Ommer 2021) and ImageBART (Esser et al. 2021) leverage a transformer to synthesize images in an auto-regressive manner while MaskGIT (Chang et al. 2022) proposes to model an image from multiple directions instead of the sequential prediction as in VQGAN. VQFR (Gu et al. 2022) designs a parallel decoder to replace the commonly used transformer for more realistic details recovery.

Unlike these methods designed for single image synthesis, our task aims to produce high-quality and temporal-consistent video portraits. Thus, we propose to learn a *personalized* codebook from a single portrait. In addition, to adapt to real scenarios, our VPNQ employs Spatio-temporal video modeling from the raw driving signals instead of the aforementioned image-modeling strategies.

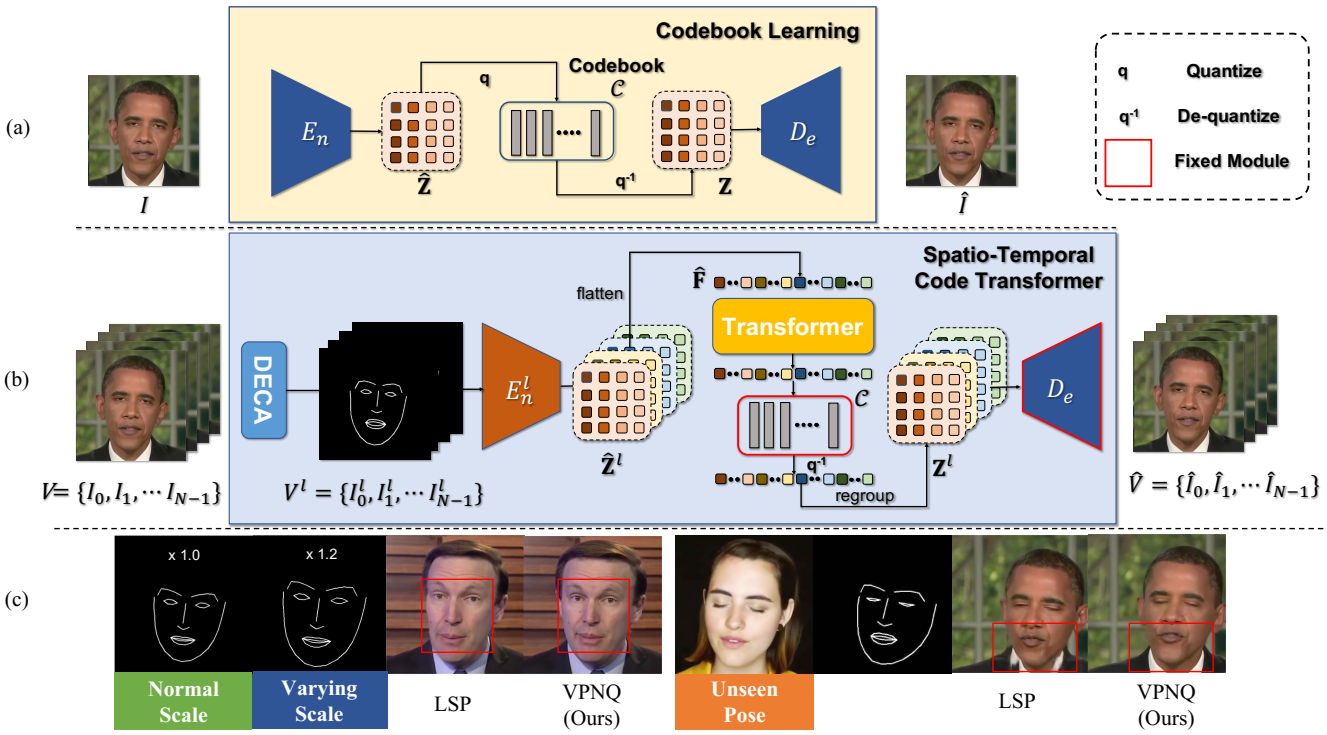

Figure 1: Overview of our VPNQ framework. (a) To achieve *personalized quantized texture* dictionary, we first learn a codebook together with an autoencoder network via self-reconstruction. (b) We propose a *Spatio-Temporal Code Transformer* based on the learned codebook and decoder for reasonable spatial composition and temporally consistent code prediction. (c) Our VPNQ can robustly recover high-fidelity video for specific portraits even under challenging scenarios (*e.g.*, varying scale and unseen pose). Please zoom in for more details.

Method

The overview of our proposed VPNQ is shown in Fig. 1 As mentioned before, in order to model a portrait with personalized quantized texture, we first build a local texture dictionary by learning a context-rich codebook and a corresponding autoencoder network. Based on the learned codebook along with the decoder, we then design a *Spatio-Temporal Transformer* module for predicting reasonable and temporally-consistent code from the input geometry information (*e.g.*, 2D landmarks). In this section, we will introduce the training procedures in these two stages detailedly.

Training and Inference Formulation. Despite the recent studies building their video portrait system based on neural- or volume-rendering, we instead choose the simplest 2D landmarks as geometry guidance in order to avoid a time-consuming training process and complex pre-processing. The training procedure is performed via self-reconstruction learning as illustrated in Fig. 1 (a). Given a video of the target portrait $V = \{I_1, \ldots, I_T\}$, we first employ the 3D parametric model DECA (Feng et al. 2021) to extract the driving landmarks $V^l = \{I_1^l, \ldots, I_T^l\}$. While in the inference stage, for a video from the driving portrait, we first use DECA to regress its morphable parameters (*i.e.*, shape, pose and expression) and replace its shape parameter with that of the target portrait to generate the driving landmarks accordingly.

Codebook Learning

Here we briefly review the image modeling procedure in VQGAN (Esser, Rombach, and Ommer 2021), which aims to represent images with compact discrete codes from a learnable codebook.

Specifically, we employ two convolutional networks with non-local attentions (Vaswani et al. 2017) as the Encoder $\boldsymbol{E}_n$ and the Decoder $\boldsymbol{D}_e$ in the quantized autoencoder network. For an arbitrary image $I \in \mathbb{R}^{H\times W\times 3}$, $\boldsymbol{E}_n$ first encodes it into a feature map $\hat{\mathbf{Z}} \in \mathbb{R}^{h\times w\times d}$, and then we *quantize* each of its spatial vectors $\hat{Z}_{(m,n)}$ by replacing it with a token $a_{(m,n)}$ corresponding to the index of its nearest neighbour in the learnable codebook $\mathcal{C} = \{c_i \in \mathbb{R}^d\}_{i=0}^{N-1}$, which is defined as:

$$a_{(m,n)} = \arg\min_i \|\hat{Z}_{(m,n)} - c_i\|_2, \tag{1}$$

where N is the codebook's size. Thus I can be represented with a set of discrete tokens $\mathcal{A} = \{a_{(1,1)}, \ldots, a_{(h,w)}\}, a_{(m,n)} \in \{0, 1, \ldots, N-1\}$. Finally, these tokens are *de-quantized* back to the feature map $\mathbf{Z} = \{c_{a_{(1,1)}}, \ldots, c_{a_{(m,n)}}\}$ through querying the codebook. By using the decoder $\boldsymbol{D_e}$, the reconstructed image can be described as $\hat{I} = \boldsymbol{D}_e(\mathbf{Z})$.

Training Objectives. We follow VQGAN (Esser, Rombach, and Ommer 2021) to leverage the L_1 loss $\mathcal{L}_{L1}$, percep-

tual loss (Wang et al. 2018) $\mathcal{L}_{per}$ and the adversarial loss $\mathcal{L}_{abv}$ (Isola et al. 2017; Wang et al. 2018) as our reconstruction supervisory. In addition, since the quantization operation is non-differentiable, the gradient copy operation in VQVAE (Oord, Li, and Vinyals 2018) is also adopted, based on the following differential loss function:

$$\mathcal{L}_{quant} = \|\boldsymbol{sg}[\hat{\mathbf{Z}}] - \mathbf{Z}\|_2 + \|\boldsymbol{sg}[\mathbf{Z}] - \hat{\mathbf{Z}}\|_2. \quad (2)$$

Here the second term is the commitment loss and $\boldsymbol{sg}[]$ denotes the "stop gradient" operation. The total training objectives for the codebook learning stage are denoted as:

$$\mathcal{L}_{cb} = \mathcal{L}_{L1} + \mathcal{L}_{per} + \mathcal{L}_{quant} + \lambda \mathcal{L}_{abv}. \quad (3)$$

More training details on tokenization can be found in (Esser, Rombach, and Ommer 2021).

Spatio-Temporal Code Transformer

We continue to discuss building a more practical framework for mapping the geometry information based on the learned codebook and the pre-trained decoder $\boldsymbol{D}_e$ in the driving signal to the decoupled local texture information. A straightforward baseline is that we can train another Encoder $\boldsymbol{E}_n^l$ with similar architecture as $\boldsymbol{E_n}$ to encode the input landmarks I^l into a feature map $\hat{\mathbf{Z}}^l \in \mathbb{R}^{h\times w\times d}$ via teacher-student learning protocol, where the feature map of the corresponding real image $\hat{\mathbf{Z}}$ encoded by the learned $\boldsymbol{E}_n$ can be used as direct supervision. However, after performing simple experiments accordingly, we observe that this baseline method fails to recover reasonable texture, especially when processing rarely seen poses. A possible explanation is that although the nearest-neighbour (NN) matching mechanism in the quantization operation may find out the nearest code for each feature vector at its local spatial position, this set of codes may not be an optimal combination for the entire face reenactment. Therefore, an alternative for this NN matching mechanism is required to alleviate this problem.

Inspired by the recent success achieved by natural language processing (Devlin et al. 2018) and transformer-based image modeling (Bao, Dong, and Wei 2021; Esser, Rombach, and Ommer 2021), we notice that the self-attention mechanism in the Transformer, which enforces *non-local interactions* among all positions on an image, is intuitively a more suitable choice for portrait modeling. Thus, we start by incorporating a Transformer module to predict more reasonable tokens for each patch instead of using the NN matching mechanism.

Spatial-part Design. Particularly, we employ a ViT-based (Dosovitskiy et al. 2021) module right behind the Encoder $\boldsymbol{E_n^l}$. In vision Transformers, the input image is first split into $P \times P$ image patches, then reshaped into a sequence of tokens as the input to the Transformer. While, in our task, the size of encoded feature map $\hat{\mathbf{Z}}^l$ from Encoder $\boldsymbol{E_n^l}$ is $h \times w \times d$. To adapt the ViT module to the previous architecture, we set $P\times P\times h\times w = H\times W$, which denotes that each spatial vector $\hat{Z}^l_{(m,n)}$ in the feature map $\hat{\mathbf{Z}}^l$ roughly matches a $P \times P$ image patch. In this paper, we set $P = 16$ by default according to the original ViT (Dosovitskiy et al. 2021).

Specifically, for an input feature map $\hat{\mathbf{Z}}^l$, we first unfold it into $h \times w$ vectors and then reform them into a flattened feature $\mathbf{F}$ before feeding it into the Transformer module:

$$\mathbf{F} = [\hat{Z}^l_{(1,1)}; \ldots; \hat{Z}^l_{(h,w)}], \hat{Z}^l_{(m,n)} \in \mathbb{R}^d \quad (4)$$

We set the layer number of ViT as 4 by default, and the procedure of *j*-th multi-head self-attention layer of the Transformer module is performed as:

$$\mathbf{F}_{j+1} = (\boldsymbol{W}_j \mathbf{V}_j) + \mathbf{F}_j, \boldsymbol{W}_j = \mathbf{Q}_j \mathbf{K}_j^\top / \sqrt{d_j}, \quad (5)$$

where $\mathbf{Q}_j$, $\mathbf{K}_j$ and $\mathbf{V}_j$ are the projected outputs of $\mathbf{F}_j$ from three separated linear layers, and d_j denotes the dimension of each head. Finally, we use another MLP network to predict the token sequence $\mathcal{A}$ and query the corresponding codes from the learned codebook $\mathcal{C}$ according to $\mathcal{A}$. After regrouping the queried codes into the quantized feature map $\mathbf{Z}^l$, we can reconstruct a high-quality image by feeding it into the pre-trained decoder $\boldsymbol{D}_e$:

$$\hat{I} = \boldsymbol{D_e}(\mathbf{Z}^l) \quad (6)$$

Temporal Training Strategy. Though the transformer module can predict reasonable tokens for each single image, we notice another issue during the exploration: the synthesized videos suffer from the temporal-inconsistency with unsatisfactory jittering on the textures.

Based on this observation, we propose a temporal strategy that involves multiple consecutive frames during training to enhance the transformer module for better temporal modeling. Specifically, for a driving video clip $V^l = \{I^l_0, I^l_1, \ldots, I^l_{T-1}\}$, we first produce a set of feature maps $\{\hat{\mathbf{Z}}^l_0, \hat{\mathbf{Z}}^l_1, \ldots, \hat{\mathbf{Z}}^l_{T-1}\}$ by using $\boldsymbol{E}_n^l$ and further flatten them into:

$$\hat{\mathbf{F}} = [\mathbf{F}^l_0, \mathbf{F}^l_1, \ldots, \mathbf{F}^l_{T-1}]^\top. \quad (7)$$

Here the flattened temporal feature $\hat{\mathbf{F}} \in \mathbb{R}^{h\cdot w\cdot T\times d}$ contains both spatial and temporal information in the whole driving sequence leading to a much larger perceptive field and thus contributes to far more temporally consistent code prediction than the single frame modeling.

Training Objectives. Since both the codebook $\mathcal{C}$ and the decoder $\boldsymbol{D}_e$ are already fixed, our goal is to precisely and consistently predict the token sequence for the clip $\hat{\mathcal{A}} = \{\hat{a}^0_{(1,1)}, \ldots, \hat{a}^0_{(h,w)}, \ldots, \hat{a}^{T-1}_{(1,1)}, \ldots, \hat{a}^{T-1}_{(h,w)}\}$, where we can define the training objective through the softmax cross-entropy loss:

$$\mathcal{L}_{code} = \sum_{\hat{a}^t_{(m,n)}\in A} -a^t_{(m,n)} \log(\text{softmax}(\hat{a}^t_{(m,n)})). \quad (8)$$

Note that $a^t_{(m,n)}$ is the ground-truth code index obtained from the corresponding real image through quantization. Particularly, to achieve robust training and faster convergence, we naturally use $\boldsymbol{E_n}$ as the teacher network for $\boldsymbol{E_n^l}$ by adding an additional L_1 loss $\mathcal{L}_{feat}$ between their encoded feature maps $\hat{\mathbf{Z}}^l$ and $\hat{\mathbf{Z}}$. Therefore, the total loss function for the Spatio-temporal transformer training is formulated as follows:

$$\mathcal{L}_{stf} = \mathcal{L}_{code} + \lambda_{feat}\mathcal{L}_{feat}. \quad (9)$$

Figure 2: Qualitative results on the HDTF dataset. We compare our VPNQ with the recent state-of-the-art methods including Facev2v (Wang, Mallya, and Liu 2021), LIA (Wang et al. 2022), NHA (Grassal et al. 2022) and LSP (Lu, Chai, and Cao 2021) under cross-reenactment setting. Please zoom in for better visualization.

λ_{feat} is set as 0.5 by default, as we expect the transformer module to release its expressive modeling power without too many constraints from the convolutional Encoder.

Experiments

Experiment Settings

Dataset and Pre-processing. We evaluate our methods on eight video sequences including five videos from the HDTF (Zhang et al. 2021) dataset, one video from AD-Nerf (Guo et al. 2021) dataset, one video from LSP (Lu, Chai, and Cao 2021) dataset and one video from Nerface (Gafni et al. 2021) dataset. Specifically, we extract each video at 60 frames per second (FPS) and crop the portrait out of the original video with the size of 512×512, so that the face can be kept at the center.

Implementation Details. All experiments are implemented on PyTorch using Adam optimizer with an initial learning rate of 5e-4 and batch size of 4. Note that, as we adopt a temporal training strategy, the 4 images in a batch are consecutive frames collected from the same video clip. The training procedure is performed in a self-reenactment manner for both two stages. For both VQGAN (Esser, Rombach, and Ommer 2021) and ViT (Dosovitskiy et al. 2021), we follow them to use their standard blocks.

Comparison Methods. We compare our VPNQ with four person-specific and two person-agnostic methods. The person-specific methods are composed of a 2D generative model method LSP (Lu, Chai, and Cao 2021) and three-volume rendering-based methods including Neural Head Avatar (NHA) (Grassal et al. 2022), AD-Nerf (Guo et al. 2021) and Nerface (Gafni et al. 2021). Regarding person-agnostic methods, we choose two latest state-of-the-art methods, Facev2v (Wang, Mallya, and Liu 2021) and LIA (Wang et al. 2022) as our counterparts.

Quantitative Evaluation

Comparison Setting. In order to quantitatively evaluate our method, we perform self-reenactment experiments on four datasets (denoted as Testset A, B C and Nerface dataset), where we randomly select 1000 frames from the test set of each subject as the driving video. Since the person-agnostic and volume rendering-based methods require larger corpus

Methods	Testset A				Testset B				Testset C			
	PSNR ↑	SSIM ↑	LPIPS ↓	F-LMD ↓	PSNR ↑	SSIM ↑	LPIPS ↓	F-LMD ↓	PSNR ↑	SSIM ↑	LPIPS ↓	F-LMD ↓
Ground Truth	N/A	1.00	0	0	N/A	1.00	0	0	N/A	1.00	0	0
Facev2v	30.16	0.70	0.16	2.52	29.58	0.67	0.14	2.47	30.99	0.78	0.09	2.45
LIA	28.79	0.71	0.22	2.51	29.41	0.66	0.18	2.82	30.22	0.74	0.13	3.03
NHA	30.49	0.67	0.28	2.22	29.79	0.68	0.17	2.14	30.84	0.72	0.24	2.23
AD-Nerf	30.30	0.67	0.18	3.60	29.45	0.64	0.20	3.49	30.41	0.67	0.15	2.84
LSP	31.26	0.76	0.08	**2.04**	30.39	0.70	0.10	2.22	31.44	0.79	0.07	**2.04**
VPNQ	**31.89**	**0.77**	**0.07**	2.07	**30.75**	**0.71**	**0.09**	**2.20**	**32.36**	**0.81**	**0.06**	2.07

Table 1: The quantitative results of on Testset A, B and C. We compare our VPNQ against recent SOTA methods (Wang, Mallya, and Liu 2021; Wang et al. 2022; Grassal et al. 2022; Guo et al. 2021; Lu, Chai, and Cao 2021) under self-reenactment setting in terms of four metrics. For LPIPS and F-LMD the lower the better, and the higher the better for other metrics.

Methods	Nerface Dataset			
	PSNR ↑	SSIM ↑	LPIPS ↓	F-LMD ↓
Ground Truth	N/A	1.000	0	0
DVP	29.10	0.74	0.32	4.42
Nerface	29.58	0.77	0.26	3.75
VPNQ	**33.17**	**0.85**	**0.07**	**2.20**

Table 2: The quantitative results on Nerface dataset.

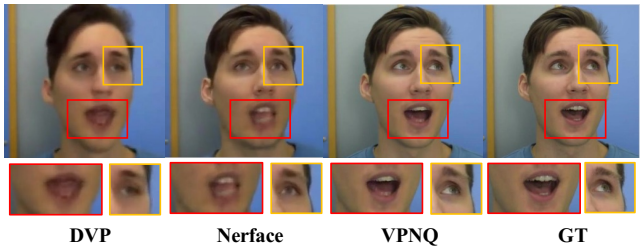

Figure 3: Qualitative results on the Nerface dataset. We provide close-up views of the areas of the mouth and eyes. Both DVP and Nerface create blurry artifacts, while our VPNQ generates clear boundaries.

training data, we crop the facial parts from generated images with the same region and then resize them to the same size for a fair comparison. Notably, in terms of person-agnostic methods Facev2v and LIA, we finetuned their released best models on each subject instead of directly using them.

Evaluation Metrics. We evaluate our method based on generation and synchronization quality. For generation quality, we follow Nerface (Gafni et al. 2021) to leverage standard metrics PSNR, SSIM, and LPIPS to measure the differences between the generated images and the ground-truth images. In terms of the synchronization quality, we use the landmark distance on the whole face (F-LMD) to measure the differences in head pose and facial expression.

Evaluation Results. The quantitative results on HDTF and Nerface datasets are summarized in Tab 1 and 2. According to the results in Tab 1, our VPNQ outperforms all the counterparts in terms of all the metrics on generation quality (*i.e.*, PSNR, SSIM and LPIPS). On the other hand, our F-LMD is slightly higher than LSP's, but much lower than those of person-agnostic methods. The results indicate that our VPNQ synthesizes high-fidelity images and achieves satisfactory synchronization.

For the results in Tab. 2, our VPNQ similarly outperforms its counterparts in all the metrics. Note that although Nerface and DVP (Kim et al. 2018) enable pose and expression manipulation, their generated results are not of satisfactory quality, including blurry in the dynamic region, which also degrades the results in the synchronization evaluation. Please also refer to Fig. 3 for visual results.

Qualitative Evaluation

We also provide qualitative evaluations and a subjective user study to demonstrate the differences between different methods better. Our qualitative evaluations are conducted under the cross-reenactment setting. Since AD-Nerf requires a template video from the same portrait as the pose input, we do not involve it in our comparison.

Evaluation Results. We show the key frames from two video clips in Fig. 2. The person-agnostic methods Facev2v and LIA achieve driving results with very similar movements to the driving video. However, even if we use the finetuned model for image synthesis, these two methods still suffer from the blurry mouth region (Please zoom in for more details). NHA and LSP either fail to recover essential textures (*e.g.*, earrings) or create blurry and incorrect textures at some local regions (*e.g.*, face boundary, and even collar) when processing unseen poses. In addition, since LSP is built upon convolutional generative models, the strong entanglement between texture and geometry also results in severe identity distortion. Our VPNQ can generate higher-quality images with more details and achieve satisfactory synchronization even under unseen scenarios.

User Study. We further conduct a user study where 15 participants are invited to evaluate our cross-reenactment results generated by our VPNQ and other comparison methods. We adopt the Mean Opinion Scores rating protocol, which requires the participants to rate the generated video portraits from three aspects: 1) Generation Quality; 2) Video Realness; 3) Synchronization. The rating is designed on a range of 1 (worst) to 5 (best).

The results are reported in Tab 3. Our VPNQ outperforms

Methods	Generation Quality	Video Realness	Synchronization
Facev2v	3.73	4.66	**4.53**
LIA	3.46	4.53	4.46
NHA	3.33	3.20	3.80
LSP	4.00	3.93	4.06
VPNQ (Ours)	**4.73**	**4.80**	4.26

Table 3: User study results based on Mean Opinion Scores. The rating is from 1 to 5; the higher the better.

	Testset C			
Methods	PSNR ↑	SSIM ↑	LPIPS ↓	F-LMD ↓
Ground Truth	N/A	1.000	0	0
Baseline	32.03	0.78	0.07	2.29
Baseline + Attention	31.88	0.76	0.07	2.40
Baseline + Spatial	**32.40**	**0.82**	0.06	2.15
Full model	32.36	0.81	**0.06**	**2.07**

Table 4: The quantitative results of ablation study.

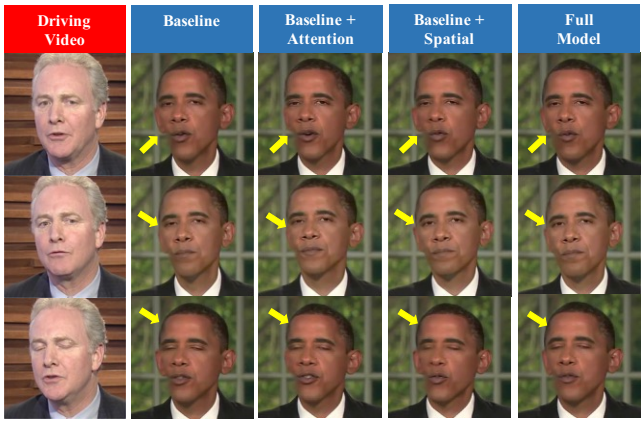

Figure 4: Qualitative results of ablation study on Testset C. Only our full model can synthesize reasonable face boundaries under very large poses.

all its counterparts in terms of generation quality and video realness thanks to the personalized quantized texture dictionary and the Spatio-temporal video modeling framework. While for synchronization, Facev2v and LIA have achieved slightly better scores than our method due to the natural movement under challenging cases. Overall, the users prefer our results in more aspects.

Ablation Study

We take Testset C as an example to perform ablation studies on three variants, which are formed by removing or replacing the modules we propose. Specifically, we first remove the whole transformer block and build a baseline with only convolutional networks (denoted as "Baseline") as mentioned in Sec 3.2. Then, to evaluate the effectiveness of our Spatio-temporal transformer, we construct another variant without adopting the temporal training strategy (denoted as "Baseline + Spatial") One may also concern if the non-local modeling can be achieved by simply inserting several self-attention layers into the variant "Baseline". Thus we design another variant denoted as "Baseline + Attention".

The quantitative results under the self-reenactment setting are reported in Tab 4. Since all the variants are trained with the local texture codebook, there is no obvious gap between their generation quality metrics. While for the synchronization metric F-LMD, the variants without the Transformer module achieve worse results.

We also provide qualitative results under the cross-reenactment setting in Fig. 4. We observe that both "Baseline" and "Baseline + Attention" fail to generate reasonable or clear boundaries for ear, face, and head parts(see the yellow arrows), which indicates that simply increasing the self-attention layers cannot work well in face reenactment. By adding spatial and Spatio-temporal transformer modules, the baseline model can progressively enable spatial composition only and Spatio-temporal modeling capability so that our full model can synthesize high-fidelity videos with the temporally-consistent face shape.

Robustness Analysis. To further evaluate the robustness of our VPNQ, we take LSP, which also relies on landmark driving, as our counterpart to conduct an additional algorithm analysis. Specifically, we provide qualitative evaluation under two challenging cases: varying scale and unseen pose. These are commonly seen distortions during the fitting of 3D models and landmarks. Here we change the scale of the driving landmarks by a ratio of 1.20 and then find a driving video with rarely seen poses. The comparison results are illustrated in Fig. 1(c), and we can observe that our VPNQ robustly generates reasonable and high-fidelity video portraits, demonstrating our robustness superiority.

Conclusion and Discussion

Conclusion. This paper presents the Video Portrait via Non-local Quantization Modeling (VPNQ) framework, which synthesizes robust and high-fidelity face reenactment results for specific portraits. We show that our VPNQ has several benefits: 1) The quantized personalized texture dictionary learned via self-reconstruction can provide more photo-realistic local textural information for the target portrait. 2)Based on the decoupled local texture dictionary and input geometry information, we can generate more robust results even under unseen cases using a non-local modeling mechanism. 3)By extending the image modeling into video modeling, we manage to synthesize temporally-consistent video via our Spatio-Temporal code Transformer.

Ethical Statement. Our method focuses on synthesizing video portraits developed for digital entertainment. However, incorrect usage of our method for malicious intentions may negatively impact social media. The Deepfake detection community has achieved impressive progress in developing robust detection algorithms to alleviate this issue. We also would like to make our efforts by sharing our generated video portraits to improve the detection algorithms to handle more complex scenarios.

References

Bao, H.; Dong, L.; and Wei, F. 2021. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*.

Blanz, V.; and Vetter, T. 1999. A morphable model for the synthesis of 3D faces. In *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*, 187–194.

Chang, H.; Zhang, H.; Jiang, L.; Liu, C.; and Freeman, W. T. 2022. MaskGIT: Masked Generative Image Transformer. *arXiv preprint arXiv:2202.04200*.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*.

Doukas, M. C.; Zafeiriou, S.; and Sharmanska, V. 2020. HeadGAN: Video-and-Audio-Driven Talking Head Synthesis. *arXiv preprint arXiv:2012.08261*.

Esser, P.; Rombach, R.; Blattmann, A.; and Ommer, B. 2021. Imagebart: Bidirectional context with multinomial diffusion for autoregressive image synthesis. *Advances in Neural Information Processing Systems*, 34.

Esser, P.; Rombach, R.; and Ommer, B. 2021. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12873–12883.

Feng, Y.; Feng, H.; Black, M. J.; and Bolkart, T. 2021. Learning an Animatable Detailed 3D Face Model from In-The-Wild Images. volume 40.

Gafni, G.; Thies, J.; Zollhofer, M.; and Nießner, M. 2021. Dynamic neural radiance fields for monocular 4d facial avatar reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8649–8658.

Grassal, P.-W.; Prinzler, M.; Leistner, T.; Rother, C.; Nießner, M.; and Thies, J. 2022. Neural head avatars from monocular RGB videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18653–18664.

Gu, Y.; Wang, X.; Xie, L.; Dong, C.; Li, G.; Shan, Y.; and Cheng, M.-M. 2022. VQFR: Blind Face Restoration with Vector-Quantized Dictionary and Parallel Decoder. In *ECCV*.

Guo, Y.; Chen, K.; Liang, S.; Liu, Y.; Bao, H.; and Zhang, J. 2021. AD-NeRF: Audio Driven Neural Radiance Fields for Talking Head Synthesis. In *IEEE/CVF International Conference on Computer Vision (ICCV)*.

Hong, W.; Ding, M.; Zheng, W.; Liu, X.; and Tang, J. 2022. CogVideo: Large-scale Pretraining for Text-to-Video Generation via Transformers. *arXiv preprint arXiv:2205.15868*.

Isola, P.; Zhu, J.-Y.; Zhou, T.; and Efros, A. A. 2017. Image-to-Image Translation with Conditional Adversarial Networks. *CVPR*.

Ji, X.; Zhou, H.; Wang, K.; Wu, Q.; Wu, W.; Xu, F.; and Cao, X. 2022. Eamm: One-shot emotional talking face via audio-based emotion-aware motion model. In *ACM SIGGRAPH 2022 Conference Proceedings*, 1–10.

Ji, X.; Zhou, H.; Wang, K.; Wu, W.; Loy, C. C.; Cao, X.; and Xu, F. 2021. Audio-driven emotional video portraits. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14080–14089.

Kim, H.; Elgharib, M.; Zollhöfer, M.; Seidel, H.-P.; Beeler, T.; Richardt, C.; and Theobalt, C. 2019. Neural style-preserving visual dubbing. *ACM Transactions on Graphics (TOG)*, 38(6): 1–13.

Kim, H.; Garrido, P.; Tewari, A.; Xu, W.; Thies, J.; Niessner, M.; Pérez, P.; Richardt, C.; Zollhöfer, M.; and Theobalt, C. 2018. Deep video portraits. *ACM Transactions on Graphics (TOG)*, 37(4): 1–14.

Lahiri, A.; Kwatra, V.; Frueh, C.; Lewis, J.; and Bregler, C. 2021. 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*, 2755–2764.

Li, L.; Wang, S.; Zhang, Z.; Ding, Y.; Zheng, Y.; Yu, X.; and Fan, C. 2021. Write-a-speaker: Text-based Emotional and Rhythmic Talking-head Generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 1911–1920.

Lu, Y.; Chai, J.; and Cao, X. 2021. Live speech portraits: real-time photorealistic talking-head animation. *ACM Transactions on Graphics (TOG)*, 40(6): 1–17.

Mildenhall, B.; Srinivasan, P. P.; Tancik, M.; Barron, J. T.; Ramamoorthi, R.; and Ng, R. 2020. Nerf: Representing scenes as neural radiance fields for view synthesis. In *European conference on computer vision*, 405–421. Springer.

Oord, A. v. d.; Li, Y.; and Vinyals, O. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Siarohin, A.; Lathuilière, S.; Tulyakov, S.; Ricci, E.; and Sebe, N. 2019. First order motion model for image animation. *Advances in Neural Information Processing Systems*, 32: 7137–7147.

Suwajanakorn, S.; Seitz, S. M.; and Kemelmacher-Shlizerman, I. 2017. Synthesizing obama: learning lip sync from audio. *ACM Transactions on Graphics (ToG)*, 36(4): 1–13.

Thies, J.; Elgharib, M.; Tewari, A.; Theobalt, C.; and Nießner, M. 2020. Neural voice puppetry: Audio-driven facial reenactment. In *European Conference on Computer Vision*, 716–731. Springer.

Van Den Oord, A.; Vinyals, O.; et al. 2017. Neural discrete representation learning. *Advances in neural information processing systems*, 30.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Wang, T.-C.; Liu, M.-Y.; Zhu, J.-Y.; Tao, A.; Kautz, J.; and Catanzaro, B. 2018. High-Resolution Image Synthesis and Semantic Manipulation with Conditional GANs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Wang, T.-C.; Mallya, A.; and Liu, M.-Y. 2021. One-shot free-view neural talking-head synthesis for video conferencing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10039–10049.

Wang, Y.; Yang, D.; Bremond, F.; and Dantcheva, A. 2022. Latent Image Animator: Learning to Animate Images via Latent Space Navigation. *arXiv preprint arXiv:2203.09043*.

Zakharov, E.; Ivakhnenko, A.; Shysheya, A.; and Lempitsky, V. 2020. Fast Bi-layer Neural Synthesis of One-Shot Realistic Head Avatars. In *European Conference on Computer Vision*, 524–540. Springer.

Zakharov, E.; Shysheya, A.; Burkov, E.; and Lempitsky, V. 2019. Few-shot adversarial learning of realistic neural talking head models. In *Proceedings of the IEEE/CVF international conference on computer vision*, 9459–9468.

Zhang, Z.; Li, L.; Ding, Y.; and Fan, C. 2021. 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*, 3661–3670.

Zheng, Y.; Abrevaya, V. F.; Bühler, M. C.; Chen, X.; Black, M. J.; and Hilliges, O. 2022. Im avatar: Implicit morphable head avatars from videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13545–13555.

Zhou, H.; Liu, Y.; Liu, Z.; Luo, P.; and Wang, X. 2019. Talking face generation by adversarially disentangled audio-visual representation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 9299–9306.

Zhou, H.; Sun, Y.; Wu, W.; Loy, C. C.; Wang, X.; and Liu, Z. 2021. Pose-controllable talking face generation by implicitly modularized audio-visual representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4176–4186.

Zhou, S.; Chan, K. C.; Li, C.; and Loy, C. C. 2022. Towards Robust Blind Face Restoration with Codebook Lookup TransFormer. *arXiv preprint arXiv:2206.11253*.