Bitrate-Controlled Diffusion for Disentangling Motion and Content in Video

Xiao Li[1†] Qi Chen[1,2,3†] Xiulian Peng[1] Kai Yu[2] Xie Chen[2,3✉] Yan Lu[1✉]
[1]Microsoft Research Asia [2]Shanghai Jiao Tong University [3]Shanghai Innovation Institute
{cq1073554383, kai.yu, chenxie95}@sjtu.edu.cn {xili11, xipe, yanlu}@microsoft.com

Abstract

We propose a novel and general framework to disentangle video data into its dynamic motion and static content components. Our proposed method is a self-supervised pipeline with less assumptions and inductive biases than previous works: it utilizes a transformer-based architecture to jointly generate flexible implicit features for frame-wise motion and clip-wise content, and incorporates a low-bitrate vector quantization as an information bottleneck to promote disentanglement and form a meaningful discrete motion space. The bitrate-controlled latent motion and content are used as conditional inputs to a denoising diffusion model to facilitate self-supervised representation learning. We validate our disentangled representation learning framework on real world talking head videos with motion transfer and auto-regressive motion generation tasks. Furthermore, we also show that our method can generalize to other type of video data, such as pixel sprites of 2D cartoon characters. Our work presents a new perspective on self-supervised learning of disentangled video representations, contributing to the broader field of video analysis and generation.

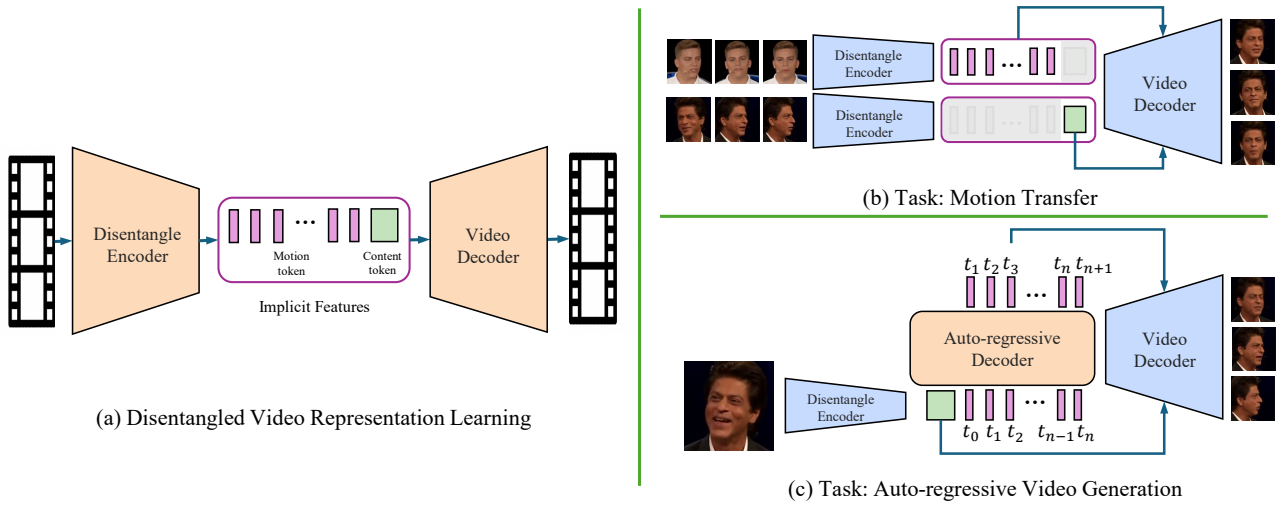
1. Introduction

Video data contains rich temporal information and intricate patterns of movement and change, offering a deeper understanding of real-world environments. Following the intrinsic information structure, a compact and robust representation of video sequences is the disentangled representation in terms of the static part (i.e. content) and the dynamic part (i.e. motion). The content part is highly semantic, encapsulating the unchanging aspects of the scene. The motion component captures the time-varying essentials of the input data, embodying the essence of change and movement within the video. Hence, the disentangled representation offers a structured space for accurately modeling, processing, and understanding the motion and the content, enabling the interpretation, the synthesis and the manipulation of the visual world in its first principles.

Despite the clear advantages of the disentangled representation of video data, the disentanglement learning task itself remains challenging. The decomposition task is intrinsically an under-determined problem, not to mention the extremely high-dimensional nature of video signals which exacerbates the computational demands as well as introduces complexities in the construction of meaningful disentanglement. Existing works either introduce additional assumptions [5, 7, 40, 42, 43, 52] or impose task-specific prior constraints [17, 23, 38, 51, 56, 62]. These additional assumptions are often overly idealistic and limit the expressiveness of disentanglement features in real scenarios, while the introduced priors constrain the target framework to specific tasks, restricting its applicability to a narrow domain.

In this paper, we propose a general and practical framework for disentangling motion and content representations in real-world video data, minimizing assumptions and task-specific inductive biases. The form of our disentangled representation is uniquely flexible - a purely implicit frame-wise motion feature and clip-wise global content latent features for a given video input sequence. Compared with other explicit or hybrid representations, implicit features have less inductive bias, significantly increasing the expressiveness of our representation. To make our disentanglement assumptions more general, we resort to the first principle and introduce an information bottleneck to regulate the information flow of implicit features during training. Inspired by recent neural codec methods [34], we achieve this by leveraging a low-bitrate codebook. Our key insight is that low-bitrate constraint acts as a general prior to enforce meaningful disentanglement, as video reconstruction under an information bottleneck shares the same goal as disentangled representations: providing a compact yet informative representation of the data [53]. We facilitate the self-supervised video reconstruction under low-bitrate with a latent denoising diffusion model. The usage of the latent diffusion model as a generative decoder for representation learning enhances the fidelity of the learned representation under information bottlenecks [31].

† Joint first authors. The major part of the work was done when Qi Chen was an intern at MSRA.
✉ Corresponding authors.



(a) Disentangled Video Representation Learning

(b) Task: Motion Transfer

(c) Task: Auto-regressive Video Generation

Figure 1. Overview of our method. (a) BCD learns compact and disentangled features to represent video data in a self-supervised manner: each video clip is represented by a clip-wise content feature and frame-wise motion feature sequences. (b) The learned disentangled video space natively supports editing tasks, e.g., swapping the content feature between two videos for motion transfer. (c) The learned representations form a structured motion space, which further enables training of auto-regressive video generation.

Putting it all together, we propose **B**itrate **C**ontrolled **D**iffusion model, **BCD** for short, a new pipeline for learning disentangled motion and content representations from video sequences (Figure 1). We demonstrate the expressiveness and the effectiveness of our proposed BCD framework by showing results on motion transfer and auto-regressive motion generation on a real world talking head video dataset [2]. Furthermore, we demonstrate the generality of our method by showing additional disentanglement results on other video data such as the LPC Sprites [48]. Our contributions can be summarized as follows:

- A novel self-supervised framework for disentangling motion and content in video with minimal assumptions and greater flexibility than prior methods.
- A low-bitrate vector quantization approach that serves as an information bottleneck to enforce meaningful disentanglement for real world videos.
- A structured and expressive latent video space enabled by a denoising diffusion model, supporting generative tasks such as motion transfer and auto-regressive video generation.

2. Related Works

Visual Representation Learning Recent years have witnessed significant advancements in image representation learning with self-supervised pre-training. The Vision Transformer (ViT) [15] paired with pretask of contrastive learning [11, 21, 25] or masked image modeling [45, 60] encodes image patches into discrete tokens and leveraging Transformer architectures [55] typically used in NLP for visual tasks. Concurrently, generative techniques like VQ-VAE [54], GANs [19], VQ-based diffusion [16, 22] or auto-regressive models [59, 61] have demonstrated remarkable capabilities in image synthesis and editing. Specifically, recent studies [31, 39] have demonstrated the potential of diffusion models in learning robust visual representations. The emergence of video diffusion models such as *Sora* [10] also significantly advances the progress in video representation. These techniques learned to compress high-dimensional visual data into lower-dimensional latent spaces while preserving essential semantic information. Our approach leverages these pioneer works, aims at representing video in a *disentangled* manner.

Disentangled Video Representation There are many works that attempts to separate motion and content from video sequences. One line of works studies the disentanglement problem from a general perspective. They introduce additional assumptions within the VAE framework such as low-dimensional features for dynamics [40], explicit dependence [52] or independence [5, 24, 43] between motion and content, or that static content can be fully modeled from a single frame in the video [7]. SKD [6] explores multi-factor disentanglement with structured Koopman autoencoders, yet they only demonstrate preliminary results on toy datasets. Our method avoids predefining motion-content correlations, allowing the aggregation of content from multiple frames, and regulate with bitrate control rather than explictly using feature with lower channels. These design choices result in a simple yet effective framework capable of handling real-world videos with complex motion and

content.

One crucial real application for video disentanglement is the generation and editing of talking head videos. Existing methods often employing explicit optical flow for motion modeling [30, 62], using specific priors such as landmarks, keypoints or 3D reconstructions [9, 17, 20, 23, 29, 38, 42, 42, 51, 56]. LIA [57] proposes to model image animation in a latent space using linear motion decomposition with a learnable orthogonal motion basis. Our goal is to explore a more general paradigm for motion-content disentanglement with minimal task-specific inductive bias. While we include talking head synthesis as an evaluation task, our approach is not tailored to solve the talking head task exclusively.

Information Bottleneck and Neural Codec The theoretical feasibility of our method is based on the information bottleneck (IB) theory. The information bottleneck (IB) provides a principled framework for learning simplified yet informative representations of data [53]. The core of IB is to retain the most relevant information of the data while discarding irrelevant details through a compression process. As noted in [1, 3], the use of IB not only yields a compact representation but also facilitates disentanglement. Our approach leverages this intrinsic property of IB as a powerful tool to disentangle video into motion and semantic content.

State-of-the-art neural video codecs [34–36, 50] have demonstrated promising results in efficient of video signals. These methods employ a strong IB with extremely high compression ratios, often relying on explicit motion vectors, such as optical flow. However, they generally do not explicitly model the disentanglement of content and motion. Recently, disentangled low-bitrate audio codecs [32] have demonstrated that a low-bitrate vector quantizer can serve as a strong information bottleneck (IB) for separating speech content from speaker identity. Inspired by these advancements in neural codecs, we apply low-bitrate constraints as an IB for video disentanglement, but with a different objective—discovering disentangled representations of motion and content within an implicit latent space.

3. Method

BCD is a self-supervised diffusion model that learns to encode video sequences into a disentangled representation of a frame-wise motion feature sequence and a sequence-wise content feature. The overall framework of BCD is illustrated in Figure 2. Following common latent diffusion approach [49], an input video sequence are first tokenized into a pre-frame latent sequence $z = \{z_t \mid t \in [1, T]\}$ with a pre-trained image VAE. Our BCD consists of a transformer encoder $\mathcal{T}(l) = (\{m_t \mid t \in [1, T]\}, c)$ that encodes the latent sequence z into its disentangled representations of per-frame motion features $m = \{m_t \mid t \in [1, T]\}$ and a global content feature c (Sec. 3.1). To promote a correct disentanglement, the per-frame motion m is bounded with a low-bitrate vector quantization bottleneck to retain only the essential motion information necessary for preserving the scene dynamics (Sec. 3.2). These disentangled representations then serve as condition inputs to guide the reconstruction of the original latent z through a diffusion model $\mathcal{D}$ (Sec. 3.3). The full model is end-to-end trained with a rate-distortion objective (Sec. 3.4).

3.1. Content and Motion Extraction

The transformer architecture [55] is pivotal to our model for the extraction and assembly of information, owing to its proven capability to facilitate information transfer across varied modalities and domains with less inductive biases [37, 42]. In our framework, we utilize a transformer encoder $\mathcal{T}$ to simultaneously aggregate information among frames and build correlations between frames in the latent space. To aggregate the content information, we prepend a fixed number of learnable queries with K tokens, i.e., $q \in \mathbb{R}^{K \times d}$, to the feature sequence $\{z_t \mid t \in [1, T]\}$ as a prefix before input into the transformer network $\mathcal{T}$. The learnable queries q are optimized over the whole training set, so that it implicitly learns to robustly aggregate information from multiple video frames. The output of the network is a prefixed latent sequence where the first K prefix token corresponding to the content feature $c \in \mathbb{R}^{K \times C_c}$, followed by the motion sequence $m \in \mathbb{R}^{T \times C_m}$. Compared to existing work which either uses image-based features with a fixed number, arbitrarily chosen reference frames, or simple pooling operators [17, 42, 51], our design can better encode content information of videos with large variations among frames (e.g., videos with different views or videos with certain details only exist in specific frames), and is flexible to the number of input frames. We implement $\mathcal{T}$ with the T5 [47] transformer equipped with relative positional encoding.

3.2. Feature Disentanglement

Separating video latents into a global content feature and a frame-wise motion sequence using only the prefixed transformer approach in Sec. 3.1, does not inherently guarantee disentangled information encoding in the corresponding features. Without constraints, the implicit features we use exhibit extremely high expressiveness, allowing arbitrary encoding of video information. This leads to two common failure modes in disentanglement [7, 52]: information preference, where the mutual information between inputs and latents is insufficient, and information leakage, where motion and content features become improperly entangled.

To ensure the disentanglement of the motion branch and content branch, we utilize Information Bottlenecks (IB) as a key component. Our key observation is that both information leakage and information preference essentially stem

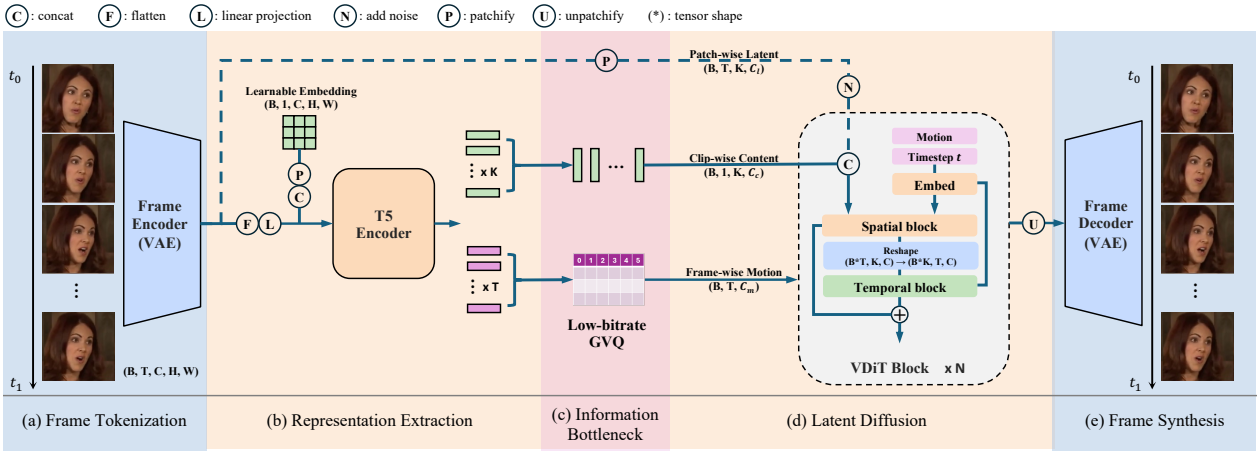


Figure 2. BCD is a diffusion model which consists of (a) a pre-trained image encoder to generate latent features for each video frame; (b) a latent transformer to extract a clip-wise content feature and frame-wise motion features; (c) a low-bitrate vector quantization module serving as information bottleneck on motion features to ensure disentanglement; (d) a latent video diffusion decoder using both content and motion features as conditions; (e) paired image decoder of (a) to synthesize final video frames.

from the improper distribution of information within the representation, failing to follow the correct disentanglement scheme. According to the information bottleneck theory, modeling a representation through an IB shares the same ultimate goal to disentanglement representations, that is, to provide a most compact yet informative representation of the data [53]. On one hand, by seeking a compact representation, the IB forces the model to discard unnecessary details, which often corresponds to the entangled factors in the data [3]. On the other hand, by preserving the most relevant information for a specific task, the model is encouraged to keep the factors that are truly meaningful and universal for the data generation process [1]. Therefore, a robust information bottleneck can not only find a compact representation, but also guide the disentanglement process by aiding in separating the underlying factors of the data. Inspired by recent work on low-bitrate neural codec [32], we implement our IB by passing the per-frame motion features $\{m_t\}$ through a bitrate-controlled vector quantization (VQ) module, detailed as follows.

Group VQ We adopt group VQ [4] to maximize the capacity upper bound of our codebooks. Specifically, the motion latent for input frame at time t, $m_t \in R^c$ is splited into N groups $\{m_t^i \in R^{c/N} \mid i \in [1, N]\}$. Each group is then individually processed by a codebook E^i with K code entries of $c_{vq} = c/N$ channels. We use a distance-gumbel-softmax layer to quantize m_t^i and sample from the codebook E^i. The sampling distribution is based on the distance between input and codes:

$$d_t^i = [\ell(m_t^i, e_1^i), \ell(m_t^i, e_2^i), \cdots, \ell(m_t^i, e_K^i)] \quad (1)$$

$$\mu_t^i = GumbelSoftmax(-\alpha \cdot d_t^i) \quad (2)$$

where $\ell(\cdot, \cdot)$ measures the L2 distance, α is a scale factor, e_j^i is the j-th entry of the i-th codebook group, and μ_t^i is the sampling distribution from E^i for input m_t^i. Quantized code $\hat{m}_t^i$ are sampled according to μ_t^i with the gumbel reparametrization trick during training and retrieved with argmax of the distribution for inference.

Bitrate Control According to Shannon's source coding theorem [12], the metric for compressibility (or compactness) of the quantized latent motion feature $\hat{m}_t^i$ is the entropy:

$$\mathcal{H}(\hat{m}_t^i | E^i) = -\sum_j P(\hat{m}_t^i = e_j^i) log(P(\hat{m}_t^i = e_j^i)) \quad (3)$$

For training, since we have the sampling distribution μ_t^i for each input, we can approximate Eq. (3) with the average sampling histogram over the training batch $\mathcal{B}$, i.e.,

$$\mu_{avg}^i = \frac{1}{|\mathcal{B}|} \sum_{\mathcal{B}} \mu_i^t \ , \ \forall \mu_i^t \in \mathcal{B} \quad (4)$$

$$\hat{\mathcal{H}}(\hat{m}_t^i | E^i) = -\sum \mu_{avg}^i log(\mu_{avg}^i) \quad (5)$$

The motion bitrate per frame from the model can be computed as $\mathcal{H}_{model} = \sum_i \mathcal{H}(\hat{m}_t^i | E^i)$. To constrain the target bitrate, we employ MSE loss between $\mathcal{H}_{model}$ and a target bitrate $\mathcal{H}_{target}$. The target bitrate $\mathcal{H}_{target}$ is a hyperparameter and we empirically set it slightly lower than the average bitrate of the video dataset (approximated from existing video codec methods).

Remarks The bitrate-controlled VQ enforces an optimal codebook that meets the average bitrate requirement for video sets while allowing the bitrate of individual videos

to remain flexible, accommodating diverse inputs. During training, the bitrate constraint is equipped with the data fidelity loss, forming a rate-distortion objective. According to rate-distortion theory, optimal compression inherently preserves expressiveness by minimizing distortion at a given bitrate. By selecting an appropriately low but nonzero bitrate, our low-bitrate controller for motion helps alleviate both information leakage and information preference. Restricting the motion bitrate prevents content from leaking into motion, as otherwise the leaked content information would cause the motion bitrate to exceed its constraint. Meanwhile, maintaining a nonzero target bitrate mitigates information preference by preventing excessive information loss in motion feature and ensuring reconstruction fidelity.

3.3. Conditional Diffusion Decoder

The BCD decoder is a conditional latent denoising diffusion model [49], which operates through a pair of forward and backward Markov chains. In the forward process, Gaussian noise ϵ_t is progressively added to erode z_t, according to a pre-defined noise schedule [33]. Conversely, the backward process performs step-by-step denoising, estimating ϵ_t to recover from z_{t-1} from z_t. The backward process is conditioned on the motion feature and content feature as guidance. We refer to [28] for closed-form equations of the diffusion process. We use DiT [44] as the backbone for our diffusion network. To ensure temporal smoothness in the generated latents, we insert additional temporal attention layers between each DiT spatial block, following [8]. We insert the motion condition by adding the motion feature to the diffusion timestep embedding, and insert the content condition by concatenate the content feature with the noisy input.

3.4. Model Training

Loss Functions Following classical rate-distortion approach [34], we train our model using both diffusion denoising loss $\mathcal{L}_d$ and the bitrate loss $\mathcal{L}_{VQ}$ with a weight factor λ:

$$\mathcal{L}_d = MSE(z, \tilde{z}) \tag{6}$$

$$\mathcal{L}_{VQ} = MSE(\mathcal{H}, \hat{\mathcal{H}}) \tag{7}$$

$$\mathcal{L} = \mathcal{L}_d + \lambda \mathcal{L}_{VQ} \tag{8}$$

We set λ to 0.04 in our experiments.

Cross-driven Strategy To further avoid collapse into trivial, entangled representations, each training video clip is evenly divided along the temporal axis into two segments with similar semantic content but distinct motion dynamics. During training, the content feature from the first segment and the motion feature from the second segment are used to reconstruct the latter segment of the video.

Implementation Details We use the pretrained image VAE from Stable Diffusion 2.0 [49] to tokenize input frame sequences. The transformer encoder has 12 layers with a hidden size of 512, a feedforward dimension of 2048 and 8 attention heads. The diffusion decoder is a DiT-B/4. The low-bitrate codebook has 64 groups, each with 32 codes. The temperature of the gumbel-softmax operator is annealed from 1.1 to 0.5 with a factor of 0.999. We adopt the EDM [33] framework for diffusion training and sampling. Please see the experiment section for training hyper-parameters and statistics.

4. Experiments

To validate our proposed BCD framework, we conduct experiments on two video datasets with distinctly different data distributions. Our main experiments are conducted on a real world video dataset consists of massive internet collected talking head videos. We choose talking heads as our primary demonstration scenario because it represents a significant real-world application of video disentanglement with in-the-wild video inputs. Additionally, fidelity and disentanglement quality can be clearly assessed through cross-identity motion transfer and video generation quality. To make our experiments more comprehensive, we also conduct a second experiment on a pixel-art style 2D character dataset, commonly used in the disentanglement literature, to provide additional disentanglement evaluation as well as to demonstrate the generalizability of our method.

4.1. Datasets

LRS3 [2] is a talking head dataset consists of face head crops from over 400 hours of TED and TEDx videos. All videos are converted to 25fps and resized to 256×256. **Sprites [48]** is a synthetic cartoon character video dataset. Each character's dynamic motion falls into one of three action categories (walking, spellcasting, or slashing) performed in three possible directions (left, front, and right). The static content includes the character's skin color, top color, pants color, and hair color, each with six possible variations. Each sequence consists of 8 frames with a resolution of 64×64.

4.2. Experiment Setup on LRS3

Hyper-parameters We set a target bitrate $H_{target} = 160$ which corresponds to 4kbps bitrate with 25fps video. The video batch size is set to 32 with 50 frames for each video clip. We train our model for 30 epochs without temporal layer, followed by 15 epochs with temporal layer, using AdamW [41] optimizer. The learning rate for first 30 epochs is 0.0001 for the decoder and 0.0002 for the encoder. For final 15 epochs, we reduce the learning rate of the encoder by half. Training on LRS3 takes $\sim$ 4–5 days on

8xA100 GPUs with 80GB memory, with an inference time of $\sim$ 30 seconds for a 50-frame video on a single A100.

Metrics For the LRS3 dataset, we adopt the FID score [26] and the cosine similarity (CSIM) computed from face recognition networks [13] to evaluate the (perceptual) video fidelity following previous talking head works. To evaluate disentanglement, we use a 3D head reconstruction method [27, 58] to fit parametric mesh coefficients and reconstruct identity mesh, motion mesh, and cross-driven ground truth mesh for reference and generated videos. We then measure the identity error, motion error, and cross-driven error by computing the mean-squared vertex distance between corresponding mesh pairs. Please refer to the supp. materials for more details.

Baselines For talking head tasks, there exists a bunch of existing works using different types of priors. We chose the following methods representing typical types of priors as our baseline: FOMM [51], MCNET [29] and LivePortrait [23] with learned keypoints and piece-wise affine flow, HyperReenact [9] with 3D parametric face priors, and LIA [57] with linear orthogonal motion basis.

4.3. Main Results on Talking Heads

Table 1. Comparisons on cross identity motion transfer. Best and runner-up methods are marked with bold and underlines.

Method	FID↓	CSIM↑	Identity error↓ $\times 10^{-1}$	Motion error↓ $\times 10^{-2}$	Cross error↓ $\times 10^{-2}$
FOMM	<u>98.5</u>	**0.76**	0.75	24.3	24.1
MCNET	98.6	**0.76**	0.85	23.9	23.6
HyperReenact	106.8	0.58	<u>0.57</u>	<u>3.94</u>	<u>4.68</u>
LIA	104.4	<u>0.71</u>	<u>0.57</u>	36.1	34.1
LivePortrait	100.3	0.69	0.66	24.6	23.7
Ours	**86.0**	0.69	**0.41**	**3.13**	**3.67**

Motion Transfer The motion transfer result is shown quantitatively in Tab. 1. Although our method does not use any human face related prior, it performs the best in terms of identity error, motion error, cross transfer error, and FID score; we also achieve a reasonable score of CSIM. We have attached more results in the supplementary materials.

To further compare our method to other baselines with different types of inductive bias, we conduct an analysis on results in the test set, and visualize generated frames in Fig. 3. FOMM [51], MCNET [29] and LivePortrait [23] share similarities with the usage of self-supervised keypoint as motion representations, which employs relative key-point locations to account for identity preservation with higher CSIM. However, they have significant larger motion (and identity) errors due to warping artifacts with large motion as well as its *same pose assumptions* between the reference frame and the first frame of the driving video. LIA [57] has the largest motion error because it exhibits a very constrained motion space with its linear combination of motion basis. Visualizations demonstrated that generated frames have good identity preservation but often completely fail to transfer motion. HyperReenact [9] leverages a 3D face prior, enabling more accurate 3D vertex positions and thus lower motion error. However, the explicit usage of a 3DMM [14] face prior makes their generated frames look like 3DMM renderings. This is reflected in its worst FID scores and can be clearly observed from the qualitative results. Compared with all these methods using different types of inductive bias, our method can synthesize high-quality motion transfer results under diverse motion inputs, even for some extreme motions while all other methods failed (Fig. 3, the first row). Please refer to our supplemental material for more video comparison results. Overall, these results clearly demonstrate that our method is able to synthesize high-fidelity videos for motion transfer via proper disentanglement of motion and content.

User Study. In addition to quantitative evaluation, we conduct a user study for subjective evaluation. The study consists of 15 groups of video reference and generated baseline videos. Users are then asked to score each generated video in terms of (1) identity preservation, (2) motion consistency and (3) visual quality. We collect responses from 18 participants in total. The result is shown in Tab. 2. As shown, we have achieved the highest score compared to other baselines, demonstrating that our disentanglement also matches human perception of video data.

Table 2. User study on identity preservation, motion consistency and visual quality. Best and runner-up methods are marked with bold and underlines.

Method	Identity Preservation ↑	Motion Consistency ↑	Visual Quality ↑
FOMM	3.34	2.63	2.84
MCNET	3.36	2.74	2.94
HyperReenact	2.97	<u>4.01</u>	<u>3.72</u>
LIA	<u>3.66</u>	2.53	3.53
Ours	**4.10**	**4.30**	**4.00**

Video Generation Our learned video space provides a structured motion representation for video generation. We demonstrate this by training an auto-regressive transformer directly on our motion spaces for the video generation task. Specifically, we train a GPT-2 [46] model on motion features generated by the BCD model on the talking head data. The content feature is inserted at the beginning as the prompt. Figure 4 visualizes some of our generated frames. We are able to generate video sequences with reasonable motion patterns based on a single-frame prompt. This experiment confirms that our motion space captures the full motion distribution rather than memorizing data points in

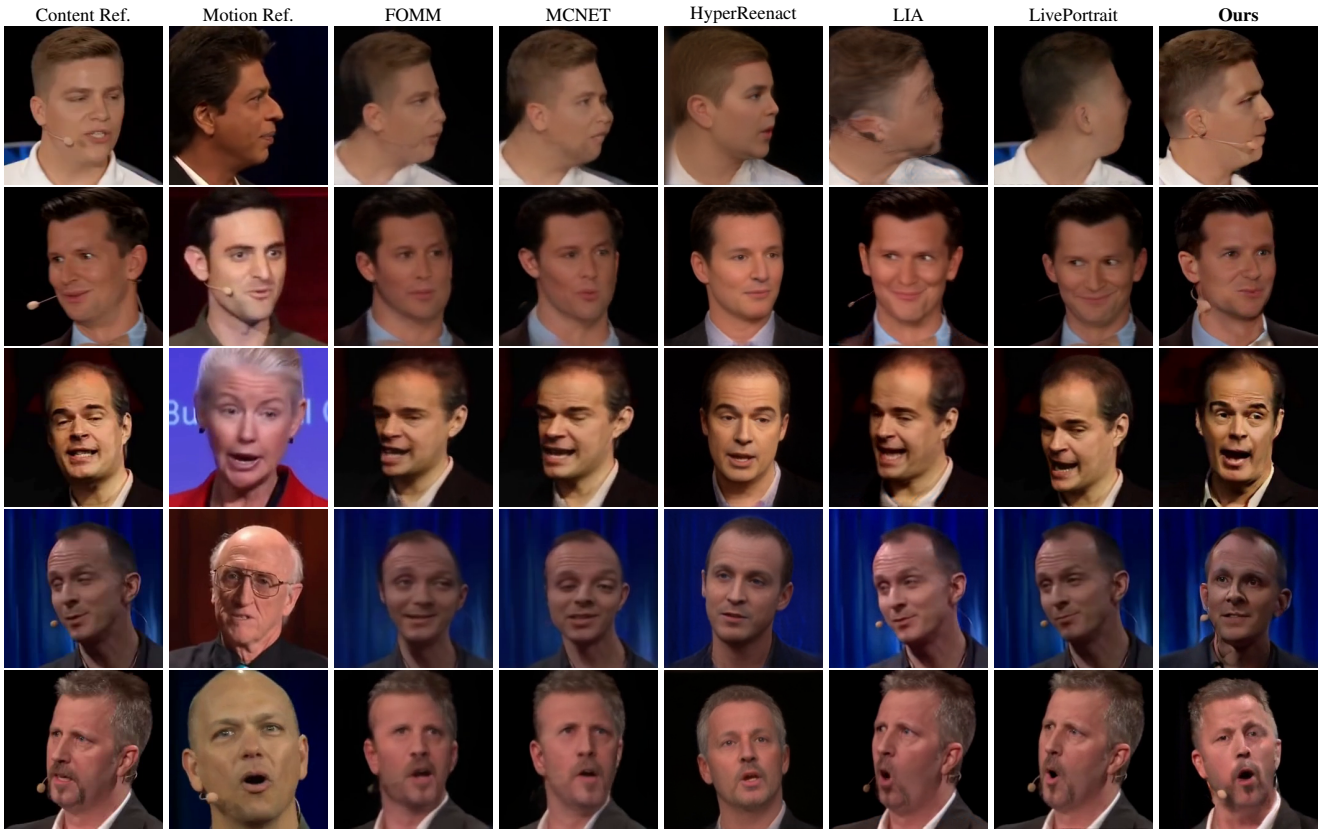


Figure 3. Motion transfer results on different input videos. The reference appearance and motion frames are shown in the first two colums.

Figure 4. Motion generation results. We use the content and motion feature of a single frame as the prompt shown on the left; five of the generated frames are shown on the right.

the training set, enabling content-preserved motion generation via sampling. We refer to the supplementary materials for the training details as well as generated video clips.

4.4. Discussions

In this section, we discuss some of our design considerations, limitations, and societal impact.

Bitrate The most important hyperparameter for BCD is the target bitrate during training. Excessive bitrate does not serve as an effective information bottleneck, resulting in diminished decoupling capabilities; conversely, a bitrate that is too low leads to insufficient transmission of information, making it struggle to synthesize high-quality videos. Both cases would lead to an increase in error for motion

Table 3. Quantitative results under different training and evaluation setup. Best and runner-up methods are marked with bold and underlines.

Target Bitrate (kbps)	FID↓	CSIM↑	Shape error↓ $\times 10^{-1}$	Motion error↓ $\times 10^{-2}$	Cross error↓ $\times 10^{-2}$
2.0	88.5	**0.71**	**0.34**	5.26	5.68
6.0	<u>87.6</u>	0.68	0.56	3.23	4.13
8.0	89.3	0.66	0.49	**3.04**	<u>3.74</u>
4.0	**86.0**	<u>0.69</u>	<u>0.41</u>	<u>3.13</u>	**3.67**
4.0 (single ref.)	87.9	<u>0.69</u>	0.47	<u>3.13</u>	3.81
4.0 (w/o. cross-driven)	120.1	0.64	0.58	41.5	40.4

transfer. We empirically set the target bitrate for motion to 4kbps, slightly lower than a recent method [18] that reports an average bitrate of approximately 5kbps for talking head videos, to promote motion disentanglement. We validate our bitrate selection in Tab. 3, which presents talking head motion transfer results across target bitrates ranging from 2 kbps (80 bits per frame) to 8 kbps (320 bits per frame). We observed that the minimum cross-driven error occurs at approximately 4kbps.

Input sequence lengths Our video disentanglement model supports a flexible number of input frames for both motion and content. As shown in Tab. 3 (row "4.0-single"), our method remains robust even with a single content frame, enabling one-shot input. Our method can be integrated with common techniques for handling transformers with long contexts, such as sliding windows. Additional long-video examples can be found in our supplemental material.

Training strategy The cross-driven strategy (Sec. 3.4) also benefits our disentanglement. The last row of Tab. 3 shows that disabling this strategy during training results in increased motion transfer error.

Societal impact Our research focuses on the self-supervised decoupling of motion and content from video. While we present results on human faces, our goal is solely to demonstrate our disentanglement capability on real videos. We have no intention of generating content for misleading or deceptive purposes. We are opposed to any behavior of creating misleading or harmful contents of real persons, and we are also interested in applying our technique for advancing forgery detection.

Limitations Our method requires substantial training data to disentangle videos with minimal task-specific inductive bias. The optimal bitrate for BCD to achieve the best disentangle can be dataset dependent. We are still observing a slight amount of video flickering even after the temporal fine-tuning. Real-world video datasets typically contain more dynamic variations and fewer static variations. This may lead to a degradation in the model's performance to capture static elements when encountering completely out-of-distribution video inputs.

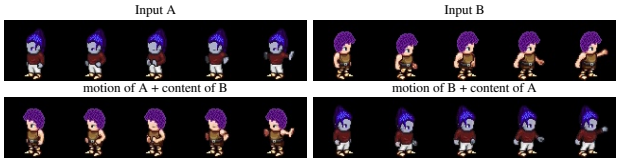


Figure 5. Motion transfer results on Sprites test set.

4.5. Results on Sprites dataset

Our proposed BCD framework is not restricted to talking head video data only, and it can potentially be applied to other video datasets that have significant different distributions. We demonstrate the generality of our BCD framework by training our model on the LPC Sprites dataset, which has been widely adopted in the disentanglement research field [5, 40, 52, 62] . The LPC Sprites dataset consists of pixel-art style, cartoon character videos in a lower resolution (4× lower than talking head videos) and fewer video clips, with a more artistic style. Hence we reduce the target bitrate H_{target} to 6 (correspond to 150bps), and directly training in pixel space instead of pre-trained image latent space. Figure 5 demonstrates motion transfer results on the LPC Sprites test dataset. Our motion transfer results demonstrate superior attribute accuracy. See the supplement for details.

5. Conclusion

We have proposed Bitrate Controlled Diffusion (BCD) for learning disentangled video representations into motion and semantic contents. BCD represents motion and contents with less inductive bias, leveraging the information bottleneck with low-bitrate vector quantization to effectively disentangle them in latent space. The training process is self-supervised with a denoising diffusion task. We have demonstrated the fidelity and the disentanglement of our method on real world talking head videos with motion transfer and generation tasks. We have also illustrated the generalization ability of our method by demonstrate results on both real world video data and synthetic cartoon video data. We believe our work offers a new perspective on disentangled video representation learning.

Future work. Future avenues include expanding our method to general video data with broader applications, exploring the interactivity of the learned motion latent space through post-hoc mapping, as well as establishing a more theoretical elaboration for the optimal bitrate to achieve disentanglement.

Acknowledgements The authors would like to thank Yue Gao for discussion, Tadas Baltrusaitis, Charlie Hewitt and Paul Mcllroy for providing 3D mesh fitting packages.

References

[1] Alessandro Achille and Stefano Soatto. Emergence of invariance and disentanglement in deep representations. *The Journal of Machine Learning Research*, 19(1):1947–1980, 2018. 3, 4

[2] Triantafyllos Afouras, Joon Son Chung, and Andrew Zisserman. Lrs3-ted: a large-scale dataset for visual speech recognition. *arXiv preprint arXiv:1809.00496*, 2018. 2, 5

[3] Alexander A Alemi, Ian Fischer, Joshua V Dillon, and Kevin Murphy. Deep variational information bottleneck. *arXiv preprint arXiv:1612.00410*, 2016. 3, 4

[4] Alexei Baevski, Steffen Schneider, and Michael Auli. vq-wav2vec: Self-supervised learning of discrete speech representations. *arXiv preprint arXiv:1910.05453*, 2019. 4

[5] Junwen Bai, Weiran Wang, and Carla P Gomes. Contrastively disentangled sequential variational autoencoder. *Advances in Neural Information Processing Systems*, 34: 10105–10118, 2021. 1, 2, 8

[6] Nimrod Berman, Ilan Naiman, and Omri Azencot. Multifactor sequential disentanglement via structured koopman autoencoders. *arXiv preprint arXiv:2303.17264*, 2023. 2

[7] Nimrod Berman, Ilan Naiman, Idan Arbiv, Gal Fadlon, and Omri Azencot. Sequential disentanglement by extracting static information from a single sequence element. *arXiv preprint arXiv:2406.18131*, 2024. 1, 2, 3

[8] Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22563–22575, 2023. 5

[9] Stella Bounareli, Christos Tzelepis, Vasileios Argyriou, Ioannis Patras, and Georgios Tzimiropoulos. Hyperreenact: One-shot reenactment via jointly learning to refine and retarget faces. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. 3, 6

[10] Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, Clarence Ng, Ricky Wang, and Aditya Ramesh. Video generation models as world simulators. 2024. 2

[11] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 2

[12] Thomas M Cover. *Elements of information theory*. John Wiley & Sons, 1999. 4

[13] Jiankang Deng, Jia Guo, Xue Niannan, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *CVPR*, 2019. 6

[14] Yu Deng, Jiaolong Yang, Sicheng Xu, Dong Chen, Yunde Jia, and Xin Tong. Accurate 3d face reconstruction with weakly-supervised learning: From single image to image set. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 0–0, 2019. 6

[15] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 2

[16] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12873–12883, 2021. 2

[17] Yue Gao, Yuan Zhou, Jinglu Wang, Xiao Li, Xiang Ming, and Yan Lu. High-fidelity and freely controllable talking head video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5609–5619, 2023. 1, 3

[18] Yue Gao, Jiahao Li, Lei Chu, and Yan Lu. Implicit motion function. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19278–19289, 2024. 8

[19] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014. 2

[20] Philip-William Grassal, Malte Prinzler, Titus Leistner, Carsten Rother, Matthias Nießner, and Justus Thies. Neural head avatars from monocular rgb videos. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18653–18664, 2022. 3

[21] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020. 2

[22] Shuyang Gu, Dong Chen, Jianmin Bao, Fang Wen, Bo Zhang, Dongdong Chen, Lu Yuan, and Baining Guo. Vector quantized diffusion model for text-to-image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10696–10706, 2022. 2

[23] Jianzhu Guo, Dingyun Zhang, Xiaoqiang Liu, Zhizhou Zhong, Yuan Zhang, Pengfei Wan, and Di Zhang. Liveportrait: Efficient portrait animation with stitching and retargeting control. *arXiv preprint arXiv:2407.03168*, 2024. 1, 3, 6

[24] Jun Han, Martin Renqiang Min, Ligong Han, Li Erran Li, and Xuan Zhang. Disentangled recurrent wasserstein autoencoder. *arXiv preprint arXiv:2101.07496*, 2021. 2

[25] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020. 2

[26] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a

two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 6

[27] Charlie Hewitt, Fatemeh Saleh, Sadegh Aliakbarian, Lohit Petikam, Shideh Rezaeifar, Louis Florentin, Zafiirah Hosenie, Thomas J Cashman, Julien Valentin, Darren Cosker, and Tadas Baltrušaitis. Look ma, no markers: holistic performance capture without the hassle. *ACM Transactions on Graphics (TOG)*, 43(6), 2024. 6

[28] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 5

[29] Fa-Ting Hong and Dan Xu. Implicit identity representation conditioned memory compensation network for talking head video generation. In *ICCV*, 2023. 3, 6

[30] Cong Huang, Jiahao Li, Bin Li, Dong Liu, and Yan Lu. Neural compression-based feature learning for video restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5872–5881, 2022. 3

[31] Drew A Hudson, Daniel Zoran, Mateusz Malinowski, Andrew K Lampinen, Andrew Jaegle, James L McClelland, Loic Matthey, Felix Hill, and Alexander Lerchner. Soda: Bottleneck diffusion models for representation learning. *arXiv preprint arXiv:2311.17901*, 2023. 1, 2

[32] Xue Jiang, Xiulian Peng, Yuan Zhang, and Yan Lu. Disentangled feature learning for real-time neural speech coding. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023. 3, 4

[33] Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. *Advances in Neural Information Processing Systems*, 35:26565–26577, 2022. 5

[34] Jiahao Li, Bin Li, and Yan Lu. Deep contextual video compression. *Advances in Neural Information Processing Systems*, 34:18114–18125, 2021. 1, 3, 5

[35] Jiahao Li, Bin Li, and Yan Lu. Hybrid spatial-temporal entropy modelling for neural video compression. In *Proceedings of the 30th ACM International Conference on Multimedia*, 2022.

[36] Jiahao Li, Bin Li, and Yan Lu. Neural video compression with diverse contexts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22616–22626, 2023. 3

[37] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023. 3

[38] Rui Li, Yiheng Zhang, Zhaofan Qiu, Ting Yao, Dong Liu, and Tao Mei. Motion-focused contrastive learning of video representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2105–2114, 2021. 1, 3

[39] Tianhong Li, Dina Katabi, and Kaiming He. Self-conditioned image generation via generating representations. *arXiv preprint arXiv:2312.03701*, 2023. 2

[40] Yingzhen Li and Stephan Mandt. Disentangled sequential autoencoder. *arXiv preprint arXiv:1803.02991*, 2018. 1, 2, 8

[41] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 5

[42] Arun Mallya, Ting-Chun Wang, and Ming-Yu Liu. Implicit warping for animation with image sets. *Advances in Neural Information Processing Systems*, 35:22438–22450, 2022. 1, 3

[43] Ilan Naiman, Nimrod Berman, and Omri Azencot. Sample and predict your latent: modality-free sequential disentanglement via contrastive estimation. In *International Conference on Machine Learning*, pages 25694–25717. PMLR, 2023. 1, 2

[44] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4195–4205, 2023. 5

[45] Zhiliang Peng, Li Dong, Hangbo Bao, Qixiang Ye, and Furu Wei. Beit v2: Masked image modeling with vector-quantized visual tokenizers. *arXiv preprint arXiv:2208.06366*, 2022. 2

[46] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019. 6

[47] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020. 3

[48] Scott E Reed, Yi Zhang, Yuting Zhang, and Honglak Lee. Deep visual analogy-making. *Advances in neural information processing systems*, 28, 2015. 2, 5

[49] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 3, 5

[50] Xihua Sheng, Jiahao Li, Bin Li, Li Li, Dong Liu, and Yan Lu. Temporal context mining for learned video compression. *IEEE Transactions on Multimedia*, 2022. 3

[51] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. First order motion model for image animation. *Advances in neural information processing systems*, 32, 2019. 1, 3, 6

[52] Mathieu Cyrille Simon, Pascal Frossard, and Christophe De Vleeschouwer. Sequential representation learning via static-dynamic conditional disentanglement. In *European Conference on Computer Vision*, pages 110–126. Springer, 2024. 1, 2, 3, 8

[53] Naftali Tishby, Fernando C Pereira, and William Bialek. The information bottleneck method. *arXiv preprint physics/0004057*, 2000. 1, 3, 4

[54] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017. 2

[55] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 2, 3

[56] 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*, pages 10039–10049, 2021. 1, 3

[57] Yaohui Wang, Di Yang, Francois Bremond, and Antitza Dantcheva. Latent image animator: Learning to animate images via latent space navigation. In *International Conference on Learning Representations*, 2021. 3, 6

[58] Erroll Wood, Tadas Baltrušaitis, Charlie Hewitt, Matthew Johnson, Jingjing Shen, Nikola Milosavljević, Daniel Wilde, Stephan Garbin, Toby Sharp, Ivan Stojiljković, et al. 3d face reconstruction with dense landmarks. In *European Conference on Computer Vision*, pages 160–177. Springer, 2022. 6

[59] Chenfei Wu, Jian Liang, Lei Ji, Fan Yang, Yuejian Fang, Daxin Jiang, and Nan Duan. Nüwa: Visual synthesis pre-training for neural visual world creation. In *European conference on computer vision*, pages 720–736. Springer, 2022. 2

[60] Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhuliang Yao, Qi Dai, and Han Hu. Simmim: A simple framework for masked image modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9653–9663, 2022. 2

[61] Wilson Yan, Yunzhi Zhang, Pieter Abbeel, and Aravind Srinivas. Videogpt: Video generation using vq-vae and transformers. *arXiv preprint arXiv:2104.10157*, 2021. 2

[62] Yizhe Zhu, Martin Renqiang Min, Asim Kadav, and Hans Peter Graf. S3vae: Self-supervised sequential vae for representation disentanglement and data generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6538–6547, 2020. 1, 3, 8