Editing Physiological Signals in Videos Using Latent Representations

Supplementary Material

1. Discussion

While our design opens up promising research opportunities and supports potential real-world applications, several limitations remain to be addressed. In the following, we discuss practical applications, outline future research potentials, and review the remaining challenges accordingly.

1.1. Applications

HR Removal Mode. Beyond targeted HR editing, our framework can also operate in a removal model. In this setting, the objective is not to replace an existing HR with a new target, but rather to eliminate the physiological trace. Technically, this can be achieved by modifying the loss functions such that both the frequency loss $\mathcal{L}_{\text{freq}}$ and wave loss $\mathcal{L}_{\text{wave}}$ are computed against pure Gaussian noise instead of a physiologically meaningful signal. At the same time, the textual conditioning prompt can be reformulated as `"Remove heart rate signal"`. In effect, the model is guided to suppress periodic rPPG components, producing videos in which HR information no longer presents while preserving visual fidelity.Table 1 reports our results demonstrating the effectiveness of the proposed removal mode. High PSNR and SSIM show that our approach preserves visual quality after signal suppression. For the physiological signals, we report both the input SNR and output SNR, along with the change in (ΔSNR), defined as the difference between the spectral power around the ground-truth HR frequency and the remaining spectrum before and after removal. The consistently negative ΔSNR values, indicate a marked suppression of periodic rPPG components, validating the effectiveness of our removal strategy.

Table 1. Quantitative results in the **HR removal mode**. PSNR and SSIM measure the visual fidelity of the output videos. Input SNR and output SNR denote the SNR of the original and modified rPPG signals, respectively. ΔSNR represents their difference, where more negative values indicate stronger suppression of the HR component.

Dataset	PSNR ↑	SSIM ↑	Input SNR	Output SNR	ΔSNR
PURE [36]	37.28	0.9532	26.82	16.37	**-10.45**
UBFC-rPPG [4]	38.53	0.9681	24.94	15.02	**-9.92**
MMPD [38]	36.80	0.9447	27.35	16.05	**-11.31**

HR in generated videos. One practical use case of our framework is that, our selected 3D VAE operates in the same latent space as modern generative backbones, potentially enabling direct integration with foundation-model pipelines. Prior studies have shown that videos generated by generative models often exhibit unrealistic HR patterns [8, 9]. To address this limitation, a Visual Language Model (VLM) can be employed to automatically determine an appropriate HR while generating the corresponding video clip in response to contextual queries (e.g., "Generate a resting scenario with a calm subject" or "Generate a running man"). The chosen HR value can then be expressed in natural language form—such as `"Heart rate 100 bpm"`—and provided as a conditioning prompt to our model. In this way, the generated videos not only reflect the intended semantic content but also exhibit physiologically consistent HR patterns, thereby supporting the creation of more realistic and precise synthetic datasets for downstream tasks, as shown in Fig. 1. It is also worth noting that our selected 3D VAE operates in the same latent space as modern generative backbones, ensuring direct compatibility with foundation-model pipelines. This design choice allows VLM-driven prompts and latent manipulations to interface seamlessly with our framework, enabling controllable and physiologically consistent video editing in future generative systems. While our current use of CLIP may appear similar to a categorical index, its formulation anticipates such extensions, positioning our method as a bridge between semantic control and physiological realism.

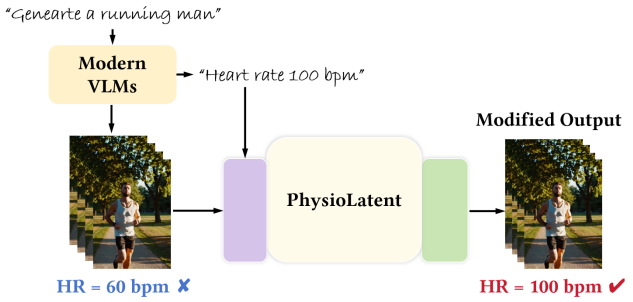

Figure 1. Illustration of a text-driven use case. A VLM generates both a video (e.g., "running man") and an associated HR prompt (e.g., `"Heart rate 100 bpm"`). The raw generated video may not exhibit a physiologically consistent HR, whereas our model ***PhysioLatent*** adjusts the rPPG signal accordingly, producing a video with both semantically correct content and physiologically accurate dynamics.

1.2. Limitations and Future Work

Lossy video reconstruction. Our approach operates in the latent space of a 3D VAE, which inherently involves a lossy

process. Even with the strongest available 3D VAE backbone, reconstruction fidelity is not perfect, and local distortions may appear in the output videos, especially in high-frequency regions, i.e., regions with edges or textures (as shown in Fig. 5), leading to a slight drop in PSNR compared to signal-processing–based baselines. One promising direction is to replace the 3D VAE backbone with emerging representations such as Gaussian Splatting [16], which have recently been explored for video embeddings [42]. Such variants could enable rPPG editing directly in the Gaussian parameter space, potentially improving reconstruction fidelity while preserving subtle physiological signals.

Lack of Heart Rate Variability (HRV) modeling. Our current design does not explicitly model HRV. The employed wave loss enforces alignment with a simple sinusoidal template at the target frequency, which facilitates HR manipulation but cannot capture the richer temporal dynamics associated with HRV. A natural extension would be to incorporate temporal generative priors such as state-space models [32] or neural ODEs [21], which can represent richer stochastic dynamics, thereby enabling explicit editing of both mean HR and its variability.

Dataset generalization. Our evaluation is currently limited to benchmark datasets, while generalization to in-the-wild conditions with challenging illumination or motion remains unexplored. Expanding the training set to more diverse scenarios would improve robustness. Large-scale pre-training with synthetic-to-real domain adaptation, or leveraging multimodal foundation models that align video with physiological signals, could significantly enhance generalization in unconstrained environments.

Incomplete frequency suppression. In some cases, the suppression of the original HR component may be incomplete: while the PSD peak shifts toward the target frequency, residual energy at the original frequency can persist, yielding multi-peak spectra and potentially confusing downstream estimators. Frequency-domain consistency losses may help achieve better suppression, ensuring that residuals are eliminated while preserving visual fidelity.

Prompt interval limitation. During training, the prompts were randomly sampled within a fixed range of 60–120 bpm at 10 bpm intervals. As a result, when unseen intermediate HR (e.g., 75 bpm) are provided as prompts, the model may fail to generate accurate outputs. Our Appendix.B includes examples using target HRs of 75 bpm, demonstrating the model's interpolation behavior between the discrete prompt values, but also indicating the limitations to be tackled in our future work. A more robust approach would be continuous prompt conditioning, where HRs are drawn from a continuous distribution rather than discrete intervals. We also provide several futuristic solutions to this issue in Appendix 3.

Potential misuse and ethical considerations. While the ability to remove or alter HR signals expands the flexibility of our framework, it also introduces the possibility of misuse. In particular, fabricated or misleading HR prompts could be presented as genuine to manipulate perceived health conditions. Such misuse might enable malicious actors to falsify physiological evidence, for instance in the context of insurance claims, medical record falsification, or biometric-based authentication. To mitigate misuse, technical safeguards such as watermarking physiological edits or cryptographic verification of unaltered signals could be developed. Alongside, ethical guidelines and responsible usage policies will be crucial for ensuring that these tools are deployed safely in medical and biometric applications.

Limited input prompt. Our current framework has been validated only using a predefined set of prompts, which ensures controlled evaluation but limits the diversity of scenarios. Extending to a broader and more flexible prompt space remains an important direction for future work, particularly in the context of generative models. For example, when integrated into a text-to-video pipeline, one may provide a single composite prompt that specifies both semantic content and physiological cues (e.g., `"Generate a running man with heart rate 120 bpm"`). Realizing such unified conditioning would require handling more natural and diverse linguistic expressions, thereby enabling tighter integration between semantic intent and physiologically consistent video synthesis—a promising future direction for our work.

2. VAE Backbone Selection and Analysis

To ensure that our framework is supported by the most reliable backbone, we perform a systematic comparison across several candidate 3D VAE models. The objective of this analysis is to demonstrate that we have carefully selected the best-performing VAE available to us. We compare 3D Causal VAE [48], TAESDV (https://github.com/madebyollin/taesd), and two variants of Video-VAE [47]. Each candidate is evaluated under a plain encoding–decoding pipeline, without any additional modification layers, so that the intrinsic reconstruction quality of the backbone can be fairly assessed.

Table 2. Comparison of different 3D VAE backbones on the PURE dataset under plain encoding–decoding (without modification layers). The 3D Causal VAE achieves the best trade-off and is therefore adopted in our framework.

3D VAE Backbone	PSNR↑	SSIM↑
3D Causal VAE	**36.90**	**0.9621**
TAESDV	22.35	0.7637
Video-VAE (4 channels)	35.38	0.9319
Video-VAE (16 channels)	35.68	0.9527

Table 2 reports the results on the PURE dataset, measuring the reconstruction quality of videos after passing through different 3D VAE backbones. Based on this comparison, we adopt the 3D Causal VAE as the backbone for our framework, since it consistently achieves superior reconstruction fidelity while maintaining robustness across different inputs.

3. Discussion on Prompt Sampling Strategy

During training, the HR prompts were randomly sampled within a fixed range of 60–120 bpm at 10 bpm intervals. While this strategy ensures sufficient coverage of typical HR values, it inherently restricts the model to a discrete set of conditions. As a result, when the model encounters unseen intermediate prompts (e.g., 75 bpm), the generated outputs may not align precisely with the desired target frequency. This limitation arises because the model has not been explicitly exposed to such in-between cases during training, thereby reducing its interpolation capability.
To illustrate this limitation, we provide an example where the input prompt is set to 75 bpm. As shown in Fig. 2, although the PSD of the generated signal exhibits a peak shift toward the intended 75 bpm frequency, the alignment is imperfect. Residual energy remains at the nearest seen training frequencies (70 bpm and 80 bpm), leading to a multi-peak spectrum. This observation highlights the interpolation gap induced by the discrete prompt sampling strategy.

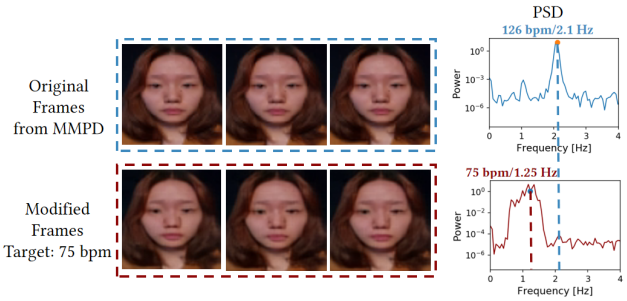

Figure 2. Qualitative example with an unseen prompt of 75 bpm. While the PSD peak is partially shifted toward 75 bpm, residual energy at neighboring frequencies remains, demonstrating limited interpolation capability. Source images are from MMPD [38].

We identify two potential directions to address this limitation in future work. First, adopting a continuous conditioning strategy, where target values are drawn from a continuous random distribution over the entire range rather than from a small set of fixed discrete intervals (e.g., 60, 70, 80), could improve generalization to unseen values. Second, introducing data augmentation in the frequency domain, such as randomized frequency shifts or adversarial perturbations, may further enhance the robustness of the learned conditioning mechanism.

4. In-the-wild Results

To complement our benchmark experiments, we also evaluated *PhysioLatent* on a small set of in-the-wild videos with challenging illumination and motion conditions from the TalkingHead1KH dataset [43]. Figure 3 presents representative examples, including both successful and unsuccessful cases. Our method performs well on videos with pronounced facial actions, where heart-rate modulation remains accurate and visually consistent. However, for sequences with large head-pose changes, the model struggles to maintain temporal coherence, leading to less precise heart-rate modification. Also, since the input videos are in-the-wild, they are out of the distribution of the training set, affecting visual quality.

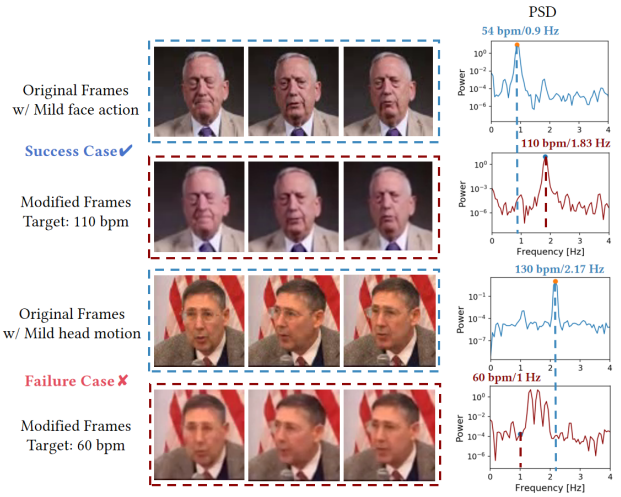

Figure 3. Representative in-the-wild results from the TalkingHead1KH dataset [43]. Our method performs well on videos with pronounced facial actions, achieving accurate and visually consistent HR modulation, but struggles with large head-pose variations, which may lead to less precise HR modification.