From Static to Dynamic Diagnostics: Boosting Medical Image Analysis via Motion-Informed Generative Videos

Wuyang Li[1], Xinyu Liu[1], Qiushi Yang[2] and Yixuan Yuan[1](✉)

[1] The Chinese University of Hong Kong, [2] City University of Hong Kong
wymanbest@outlook.com, yxyuan@ee.cuhk.edu.hk

Abstract. In the field of intelligent healthcare, the accessibility of medical data is severely constrained by privacy concerns, high costs, and limited patient cases, significantly hindering automated clinical assistance. Though previous efforts have been made to synthesize medical images via generative models, they are limited to *static* imagery that fails to capture the *dynamic* motions in clinical practice, such as contractile patterns of organ walls, leading to vulnerable prediction in diagnostics. To tackle this issue, we propose a holistic paradigm, VidMotion, to boost medical image analysis with generative medical videos, representing the first exploration in this field. VidMotion consists of a Motion-guided Unbiased Enhancement (MUE) to augment static images into dynamic videos at the data level and a Motion-aware Collaborative Learning (MCL) module to learn with images and generated videos jointly at the model level. Specifically, MUE first transforms medical images into generative videos enriched with diverse clinical motions, which are guided by image-to-video generative foundation models. Then, to avoid the potential clinical bias caused by the imbalanced generative videos, we design an unbiased sampling strategy informed by the class distribution prior statistically, thereby extracting high-quality video frames. In MCL, we perform joint learning with the image and video representation, including a video-to-image distillation and image-to-image consistency, to fully capture the intrinsic motion semantics for motion-informed diagnosis. We validate our method on extensive semi-supervised learning benchmarks and justify that VidMotion is highly effective and efficient, outperforming state-of-the-art approaches significantly. The code is available at https://github.com/CUHK-AIM-Group/VidMotion.

Keywords: Medical image analysis · Generative medical videos · Motion-informed diagnosis · Semi-supervised learning

1 Introduction

The explosion of large models [35,2] has profoundly impacted our daily life, primarily driven by the extensive data availability [30]. However, acquiring adequate medical images is particularly challenging [5,20,15,23] in the medical field

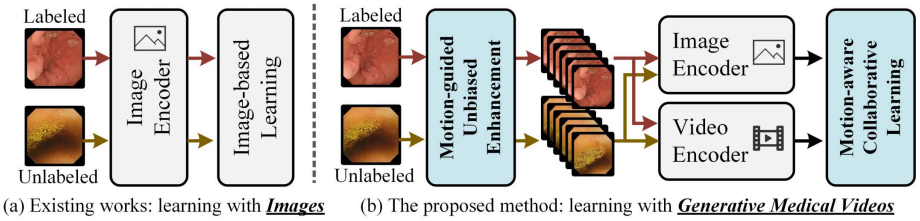

Fig. 1. Illustration of (a) existing pipelines and (b) the proposed VidMotion framework.

due to privacy concerns, high costs, and limited patient cases [34,9], posing significant hurdles to developing diagnosis systems [32,26,22,14,27] and the medical community [37,7,3,24].

To break through this barrier, considerable progress [32,28,1] has been made to scale up data with medical image synthesis, which can broaden the diversity of datasets with generative models [32,12,2]. The authors in [28] employ diffusion models to achieve style translation, effectively bridging medical domain gaps. Allmendinger *et al.* [1] and Frisch *et al.* [5] tackle the imbalanced medical imaging generation by delving into rare diseases and surgical operations. Some works [8] focus on synthesizing tumor cases to improve tumor detection. Recently, there have been notable strides in generating various data types, such as lung CT, retina, and pathological images [32], enriching the data resource significantly.

Despite the great progress, existing works predominantly focus on synthesizing ***static*** images, which fail to capture the ***dynamic*** nature of clinical environments, such as surgical movement [25] and blood flow, undermining the robustness and accuracy of clinical practice. To this end, it is natural to draw inspiration from medical videos enriched with motion-based semantics. Compared with static imaging, the dynamic nature of videos can model richer and more critical cues, such as subtle movements and the progression of symptoms over time, which are essential for accurate disease identification and monitoring [33]. Recently, some works [6,2,11] have just emerged to explore video generation beyond individual images. AnimateDiff [6] and Stable Video Diffusion [2] design powerful motion modules to capture temporal dependence, bringing the generation to a new level.

Hence, as the first exploration, this paper aims to boost medical image analysis via generative medical videos, thereby enabling the perception of clinical motions. However, there are two challenges in achieving such a reliable motion-informed diagnostic. First, directly enhancing medical images for all classes equally with generative videos will exacerbate the class imbalance issue [5], because head classes tend to yield imbalanced video generation, leading to biased diagnoses. Second, as shown in Fig. 1(a), current methods mainly learn with static images, failing to capture video-based dynamics. Compared with static images, the dynamic motions captured in videos, *e.g.*, subtle movements of mucosal surfaces, contractile patterns of organ walls, and the dynamic interaction between instruments and tissues, provide invaluable information in clinical as-

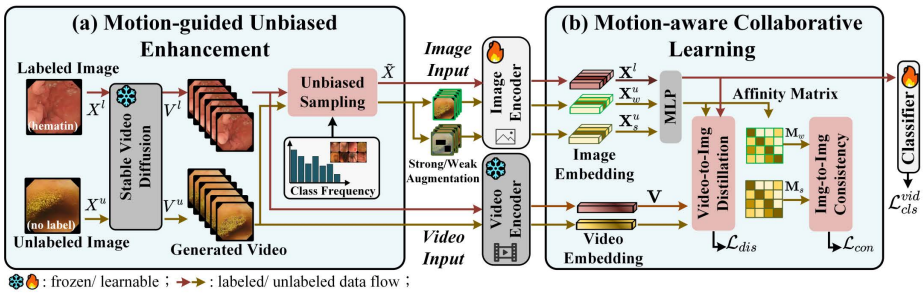

Fig. 2. Overview of our VidMotion. (a) MUE generates medical videos and conducts unbiased sampling. (b) MCL learns motion semantics with images and videos jointly.

sessment [25]. Thus, a meticulous understanding of video-based motion patterns is imperative for enhancing medical image analysis and therapeutic strategies.

To tackle the challenge, we propose a holistic framework, VidMotion, to boost medical image analysis with video-driven motion, as shown in Fig. 1(b). VidMotion consists of Motion-guided Unbiased Enhancement (MUE) to augment static images with generative medical videos unbiasedly and Motion-aware Collaborative Learning (MCL) to capture the video dynamics. Specifically, MUE enhances medical images into short videos enriched with diverse clinical motions and conducts unbiased sampling to gather reliable frames statistically. Then, MCL deploys video-to-image distillation and image-to-image consistency to capture the motion-based semantics, thereby improving the diagnosis with video dynamics. Considering that the generated videos can boost various types of data, we preliminarily evaluate VidMotion with the semi-supervised learning (SSL) diagnosis benchmark, *i.e.*, a clinically practical setting using labeled data and unlabeled data, to thoroughly assess the capacity of both supervised and unsupervised scenarios [29,21,16]. Extensive experiments verify that VidMotion significantly surpasses SOTA methods. Besides methodology contributions, our synthesized high-quality videos can contribute to medical research greatly.

2 Methodology

In SSL, we have labeled data $X^l = \{(x_i^l, y_i^l)\}_{i=1}^{B_l}$ and unlabeled data $X^u = \{(x_i^u)\}_{i=1}^{B_u}$ to train the model, where B_l and B_u denotes the corresponding batch size. The overview of the proposed VidMotion is shown in Fig. 2. Given the labeled and unlabeled data $\{X^l, X^u\}$, MUE (Fig. 2(a)) uses leverages the frozen Stable Video Diffusion [2] model[1] to generate N video frames $\{V^l, V^u\}$ for each image, and then conducts unbiased sampling to collect a sub-set of video frames $\tilde{X}^{l/u} \subset V^{l/u}$. Next, the sampled video frames $\tilde{V}^{l/u}$ and complete video streams $V^{l/u}$ are sent to a learnable image encoder and frozen video encoder to generate

[1] The used SVD is pre-trained on large-scale online videos with superior generalization capacity, spreading natural and medical domains.

image $\tilde{\mathbf{X}}^{l/u} = \{\mathbf{x}_{i,k}\}_{i=1;k=1}^{B,K}$ and video embedding $\mathbf{V}^{l/u} = \{\mathbf{v}_i\}_{i=1}^{B}$, respectively, where K is the number of sampled frames. Finally, MCL (Fig. 2(b)) distills motion semantics from videos to boost the image representation.

2.1 Motion-guided Unbiased Enhancement

Different from existing works [31,13] only using static images, we aim to synthesize medical videos with motion semantics, which is crucial for enhancing model robustness against clinical motions, *e.g.*, the instrument movements. To this end, we leverage Stable Video Diffusion [2], an advanced video generative model, to generate videos from medical images. Specifically, the generation process is formulated with a diffusion process in a Markov chain, which can generate video data v_0 from the noise $v_T \sim \mathcal{N}(0, 1)$ via a T-step denoising process guided by a specific condition. In this process, we use the labeled and unlabeled data as the diffusion condition to guide the generation, enabling the videos to have similar content with images. The generation process is as follows given image x,

$$p_\theta(v_{0:T}|x, \gamma) = p(v_T) \prod_{t=1}^{T} p_\phi(v_{t-1} \mid v_t, x, \gamma), \tag{1}$$

where ϕ is the pre-trained Stable Video Diffusion model [2], $p_\phi()$ indicates the estimated conditional distribution for generated medical videos, $\gamma \in [0, 255]$ is a constant controlling the motion intensity of generated videos. Then, for each image batch $X^{l/u} = \{x_i^{l/u}\}_{i=1}^{B}$, a set of synthesized videos $V^{l/u} = \{x_{i=1}^{l/u}\}_{i=1}^{B}$ are obtained to model diverse motions, where $v_i^{l/u} = \{(v_{i,1}^{l/u}, x_{i,2}^{l/u}, x_{i,3}^{l/u}, ..., x_{i,N}^{l/u})\}$ indicates the video frames generated by image $x_i^{l/u}$, and N is the number of frames. In our experiments, we find that generated videos adhere to satisfactory physical rationality, effectively simulating various motions in clinical practice, *e.g.*, spatial translation, liquid flow, shake blur, etc. (see Fig. 3 for visualization)

Unbiased Sampling. As medical data significantly suffers from class imbalance, the rare cases are overshadowed by an abundance of common cases, detrimentally influencing model learning and diagnosis accuracy. This issue becomes more pronounced [5] when scaling up data with videos since the prevalent classes yield more video frames with larger medical diversity. To avoid such negative influence, we deploy a simple yet effective mechanism to conduct unbiased sampling on the generated video frames according to the class distribution prior. Specifically, given C classes with N_c labeled samples for class c, we collect a subset of video frames $\tilde{X}^{l/u}$ with the guidance of the class frequency:

$$\tilde{X}^{l/u} = \text{RandomSample}(V^{l/u}, \lceil \alpha \cdot |V^{l/u}| \rceil), \text{ where } \alpha = \frac{\frac{1}{N_c}}{\sum_{j=1}^{C} \frac{1}{N_j}}, \tag{2}$$

and $V = \{v_i\}$ is all synthesized videos. Thus, the proposed unbiased sampling enhances the rare classes by collecting more frame images, promoting a balanced distribution for diagnostic fairness without clinical bias.

2.2 Motion-aware Collaborative Learning

With the generated videos $V^{l/u}$ and the sampled image frames $\tilde{X}^{l/u}$, we aim to conduct collaborative learning between the image and video modalities. Considering that the video contains rich temporal information and motion cues, the model is encouraged to generate motion-robust predictions for clinical practice. Specifically, the sampled video frames $\tilde{X}^{l/u}$ with $|\tilde{X}^{l/u}| = K_{l/u}$ are sent to the image encoder to generate image embedding $\mathbf{X}$, where the labeled data yields $\mathbf{X}^l \in \mathbb{R}^{B_l \times K_l \times D'}$ and the unlabeled data is conducted strong/weak augmentation [31] to yield $\mathbf{X}^u_s$ and $\mathbf{X}^u_w$, where $\mathbf{X}^u_{s/w} \in \mathbb{R}^{B_u \times K_u \times D'}$. At the same time, we send generated videos $V^{l/u}$ to a pre-trained video encoder [35] to encode temporal-aware knowledge, yielding the video embedding $\mathbf{V}^{l/u} \in \mathbb{R}^{B_{l/u} \times D}$.

Video-to-Image Distillation. To extract the inherent motion cues at the temporal axis, we propose embedding distillation to transfer the video semantics to the image counterpart, enabling motion perception in the image branch. To this end, given the video embedding $\mathbf{V}$ and the image embedding $\mathbf{X}$, we first deploy an MLP projection layer on the image embedding to scale up the dimension for more representative space. As we adopt the same operations for labeled and unlabeled samples, we do not write the superscripts (l/u) of the embedding for mathematical clarity. Then, we distill the motion-aware cues from the video embedding to associated image frames with L_1 loss, which is denoted as follows,

$$L_{dis} = \frac{1}{B \times K \times D} \sum_{b=1}^{B} \sum_{k=1}^{K} \sum_{d=1}^{D} \left| \frac{\mathrm{MLP}(\mathbf{X})_{[b,k,d]}}{|\mathrm{MLP}(\mathbf{X})|} - \frac{\mathbf{V}_{[b,d]}}{|\mathbf{V}|} \right|_1, \tag{3}$$

where $\mathrm{MLP}(\mathbf{X}) = \mathbf{W}^{(2)}(\mathrm{ReLU}(\mathbf{W}^{(1)}\mathbf{X} + \mathbf{b}^{(1)}) + \mathbf{b}^{(2)}$ with learnable weight $\mathbf{W}$ and bias $\mathbf{b}$. This cross-modality distillation can transfer the temporary semantics to the image model, thereby ensuring the motion robustness.

Image-to-Image Consistency. To harness the abundant inter-frame dependencies for reliable model recognition, we further enhance cross-image consistency within the imaging modality. Thus, we enable the model to leverage the rich temporal knowledge within video sequences. Specifically, given the image embedding of strong/weak augmented unlabeled data $\mathbf{X}^u_{s/w} \in \mathbb{R}^{B_u \times K_u \times D'}$, we reuse the former MLP projection layers to generate embedding and then calculate the pair-wise cosine similarity to generate affinity matrix $\mathbf{M}^u_{s/w} = \frac{\mathbf{X}^u_{s/w} \cdot (\mathbf{X}^u_{s/w})^T}{\|\mathbf{X}^u_{s/w}\|_2 \cdot \|\mathbf{X}^u_{s/w}\|_2}$, $\mathbf{M}^u_{s/w} \in \mathbb{R}^{B_u \times K_u \times K_u}$. Then, we encourage the consistency between the affinity matrix obtained from the strong and weak augmented samples, which can be expressed as the following loss function,

$$L_{con} = \frac{1}{B_u \times K_u \times K_u} \sum_{i=1}^{B_u} \sum_{j=1}^{K_u} |\mathbf{M}^u_{s\,[b,k]} - \mathbf{M}^u_{w\,[b,k]}|_2. \tag{4}$$

Different from existing works [29,21,36] that typically process images independently, in the proposed method, the relation within each video frame pair can be thoroughly enhanced via the consistency loss [10], boosting the image model with long-distance dependence [19,17,18] among different video frames.

Table 1. Comparison with SOTA methods on Kvasir-Capsule and ISIC 2018 datasets.

Method	5%			10%			20%			40%		
	MAP	MAR	AUC	MAP	MAR	AUC	MAP	MAR	AUC	MAP	MAR	AUC
Kvasir-Capsule: Endoscopic Scene												
FixMatch [31]	66.77	56.84	76.83	69.36	58.59	78.04	80.75	68.88	83.39	85.87	76.51	87.54
CoMatch [13]	68.11	63.22	80.44	73.80	65.19	81.71	82.74	71.30	84.74	86.07	79.88	89.15
SimMatch [13]	67.25	65.69	81.77	70.43	71.37	84.56	82.24	70.44	84.58	86.81	81.25	89.95
TEAR [29]	67.46	65.71	81.65	69.83	72.36	82.23	82.35	73.28	85.99	87.78	80.94	90.02
ACPL [21]	70.17	67.21	81.97	74.73	66.46	82.33	83.42	74.45	86.52	87.41	82.76	90.85
SimMatchV2 [13]	70.96	65.99	81.78	74.91	75.29	84.20	84.34	75.08	86.79	87.91	85.31	92.11
VidMotion	**73.55**	**69.96**	**83.75**	**78.28**	**77.57**	**87.91**	**86.05**	**79.89**	**89.34**	**91.21**	**86.41**	**92.70**
ISIC 2018 Skin Lesion: Dermoscopic Scene												
FixMatch [31]	37.61	25.49	57.47	38.04	30.27	60.60	43.78	37.80	64.73	49.32	41.06	66.75
CoMatch [13]	39.04	25.95	57.84	39.77	29.45	60.22	45.51	37.84	65.15	50.29	41.29	67.27
SimMatch [13]	39.25	26.09	58.71	41.05	30.00	60.65	44.87	39.49	65.81	51.77	42.64	67.21
TEAR [29]	40.90	25.61	57.95	42.00	30.60	61.34	45.20	39.71	65.73	50.55	41.73	67.24
ACPL [21]	41.67	25.07	57.44	43.42	32.24	62.14	45.29	38.06	65.19	51.76	42.49	68.11
SimMatchV2 [13]	41.50	27.61	58.90	43.82	33.05	62.42	46.38	38.14	65.31	51.72	43.92	68.43
VidMotion	**44.25**	**28.16**	**59.76**	**45.46**	**34.55**	**63.24**	**47.14**	**42.25**	**67.37**	**54.19**	**46.39**	**69.71**

2.3 Training and Inference

In the training stage of VidMotion, we implement the following loss function:

$$\mathcal{L} = \lambda_1 L_{dis} + \lambda_2 L_{con} + L_{cls}^{vid} + L_{base}, \tag{5}$$

where L_{dis} is the video-to-image distillation loss, L_{con} is the image-to-image consistency, L_{cls}^{vid} is the standard classification loss for sampled video frames, and L_{base} can be deployed as any SSL baseline. Note that generated videos can maintain semantic consistency with reference images to a certain degree since the disease area may move out of the frame in some cases. Nonetheless, we assign an image-consistent label to the generated video frames. In the inference stage, we only implement the image encoder and classifier for the diagnosis.

3 Experiments

3.1 Experimental Setup

Datasets. We evaluate our methods on two public benchmarks with extensive settings. **(1) Kvasir-Capsule.** KC is a real-world endoscopic dataset containing 47,238 images with 14 challenging clinic classes. We follow existing works [29] to randomly collect the subset for the model training and test for fair comparison. **(2) ISIC 2018.** ISIC 2018 is a real-world skin lesion dataset [4], which consists of 10,015 dermoscopy images. ISIC contains seven kinds of different skin lesions[2], which is a more challenging dataset with the intrinsic class-imbalanced issue. Different from existing work [29] relying on the class-balanced data splitting, we

[2] For dermatology images, generated videos simulate rational camera movements, e.g., translation and zoom, which are crucial for performance gains.

Table 2. Ablation study results on Kvasir-Capsule and ISIC 2018 datasets.

Setting			Kvasir-Capsule						ISIC 2018 Skin Lesion					
MUE	MCL		5%			40%			5%			40%		
	V2I	I2I	MAP	MAR	AUC	MAP	MAR	AUC	MAP	MAR	AUC	MAP	MAR	AUC
×	×	×	68.11	63.22	80.44	86.07	79.88	89.15	39.04	25.95	57.84	50.29	41.29	67.27
✓	×	×	71.87	65.72	81.48	88.77	82.33	90.54	43.22	26.62	58.71	53.10	44.10	68.63
✓	✓	×	72.51	68.40	83.06	91.03	84.35	91.02	44.02	27.23	59.01	53.14	45.23	69.02
✓	✓	✓	**73.55**	**69.96**	**83.75**	**91.21**	**86.41**	**92.70**	**44.25**	**28.16**	**59.76**	**54.19**	**46.39**	**69.71**

Table 3. Sensitivity on loss weight λ.

λ_1	λ_2	MAP	MAR	AUC
0.1	1.0	73.55	**69.96**	**83.75**
0.2	1.0	**74.01**	69.31	82.97
0.1	2.0	73.12	69.42	82.23
0.05	1.0	72.96	68.88	82.12
0.1	0.5	73.24	69.02	83.01

Table 4. Sensitivity on motion γ.

γ	MAP	MAR	AUC
55	72.11	68.33	82.07
105	72.48	69.02	83.03
155	73.03	69.11	83.38
205	73.21	69.33	83.42
255	**73.55**	**69.96**	**83.75**

conduct four different SSL settings with 5%, 10%, 20%, and 40% label regimes according to the real class distribution for more clinical rationality.

Evluation Metrics. To thoroughly evaluate SSL in real-world situations, we use three evaluation metrics for strict comparison, including Macro-Average Precision (MAP), Macro-Average Recall (MAR), and multi-class Area Under Curve (AUC), where MAP and MAR can better evaluate imbalanced medical scenarios, and AUC can better analyze the general performance in the balanced situation.

Implementation Details. We follow [31,13] to implement all methods on WideResNet-22 image encoder and deploy the pretrained CLIP-ViP [35] video encoder. For video generation, we use SVD-XT [2] to generate $N = 25$ video frames[3] for each medical image with $T = 25$ using one NVIDIA A100 GPU. The motion intensity γ is set to 255 to maximize the motion diversity. Considering the computation cost [2], we use 5% images to generate videos as hold-out experiments, which can be further improved with larger ratios. We train learnable models with 100 epochs with SGD optimizer, the learning rate of 1×10^{-2}, a momentum of 0.9, and a weight decay of 5×10^{-4}. Experiments are performed on NVIDIA 2080 Ti GPUs with $N_l = 12$ and $N_u = 84$. The strong/weak augmentations are consistent with baseline [13] for fair comparison. The loss weights λ_1 and λ_2 in Eq. 5 are empirically set as 0.1 and 1.0, respectively.

3.2 Quantitative Study

Comparison with State-of-the-Arts. As shown in Tab. 1, we compare the proposed VidMotion with state-of-the-art SSL methods with different label regimes. Compared with the most advanced SSL method SimMatchV2 [38], our method achieves consistent and noticeable gains on all evaluation matrices, which performs 2.59%, 3.37%, 1.71%, and 3.3% MAP gains, and gives 1.97%, 3.71%,

[3] The 25 frames consists of 1 given image and 4 seconds of 6-FPS video.

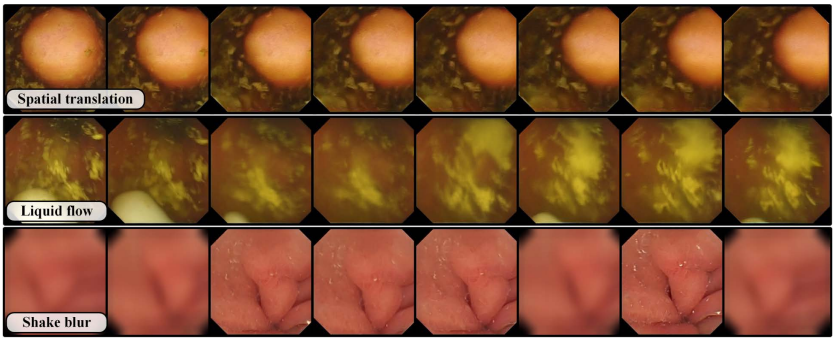

Fig. 3. Illustrating the generative video frames. Generated videos can successfully simulate diverse clinical motions, *e.g.*, spatial translation, liquid flow, and shake bur.

2.06%, and 0.69% AUC improvements. This indicates that our method is highly effective and robust to the data distribution with great generalization capacity. In comparison with the advanced SSL method in the field of medical imaging, the proposed surpasses TEAR [29] and ACPL [21] with 2.10% and 1.97% AUC (5%), respectively, showing our strong capacity under data-efficient learning.

Ablation Analysis. We report detailed ablation analysis in Tab. 2 on each designed component, evaluated on two benchmarks under two different label regimes. Compared with the baseline model with 68.11%, 86.07% 39.04%, and 50.29% MAP, introducing video-enhanced data for training (MUE) gives significant performance gains with 71.87%, 88.77%, 43.22%. and 53.10% MAP, verifying the critical motion-based semantics. Then, after introducing MCL with V2I and I2I, we can observe noticeable performance improvements with 73.55%, 91.21%, 44.25%, and 54.19% MAP, which surpasses the baseline model with significant 5.44%, 5.41%, 5.21%, and 3.90% MAP improvements, revealing the superior effectiveness of the proposed collaborative learning paradigm.

Sensitivity Analysis. To further analyze our VidMotion, we conduct a detailed sensitivity analysis on the core hyper-parameters. In Tab. 3, if we decrease the loss weight with $\lambda_1 = 0.05$ and $\lambda_2 = 0.5$, there is a small performance decrease (-1.05% and -0.77% MAP) compared with our optimal setting, indicating the effectiveness of our design. In Tab. 4, our method is robust to the motion intensity and gives slight gains when we enlarge the γ due to more diverse motion types.

3.3 Qualitative Study

As shown in Fig. 3, we present the video frames generated by the images in three different classes. The left-most image in each row represents the reference image for the image-to-video generation. We are impressed that the generated videos not only adhere to the laws of physical motion but also successfully simulate diverse movements in clinical environments, including spatial translations, fluid dynamics, and shaking bur, which is robust to diverse classes.

3.4 Conclusion

This paper proposes a holistic framework named VidMotion to boost medical image analysis with generative medical videos, which breaks through the static diagnosis in existing works by learning with dynamic videos. VidMotion consists of a Motion-guided Unbiased Enhancement module to augment medical images into motion-informed videos at the data level. Besides, it designs a Motion-aware Collaborative Learning module to encourage the joint learning of image and video modalities. Extensive experiments verify that our method is both highly effective and efficient, which surpasses SOTA methods by a large margin.

Acknowledgement. This work was supported by Innovation and Technology Commission- Innovation and Technology Fund ITS/229/22 and Hong Kong Research Grants Council (RGC) General Research Fund 11211221, 14204321.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Allmendinger, S., Hemmer, P., Queisner, M., Sauer, I., Müller, L., Jakubik, J., Vössing, M., Kühl, N.: Navigating the synthetic realm: Harnessing diffusion-based models for laparoscopic text-to-image generation. ArXiv:2312.03043 (2023)
2. Blattmann, A., Dockhorn, T., Kulal, S., Mendelevitch, D., Kilian, M., Lorenz, D., Levi, Y., English, Z., Voleti, V., Letts, A., et al.: Stable video diffusion: Scaling latent video diffusion models to large datasets. ArXiv:2311.15127 (2023)
3. Chen, Z., Li, W., Xing, X., Yuan, Y.: Medical federated learning with joint graph purification for noisy label learning. Medical Image Analysis **90**, 102976 (2023)
4. Codella, N., Rotemberg, V., Tschandl, P., Celebi, M.E., Dusza, S., Gutman, D., Helba, B., Kalloo, A., Liopyris, K., Marchetti, M., et al.: Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic). ArXiv:1902.03368 (2019)
5. Frisch, Y., Fuchs, M., Sanner, A., Ucar, F.A., Frenzel, M., Wasielica-Poslednik, J., Gericke, A., Wagner, F.M., Dratsch, T., Mukhopadhyay, A.: Synthesising rare cataract surgery samples with guided diffusion models. In: MICCAI (2023)
6. Guo, Y., Yang, C., Rao, A., Liang, Z., Wang, Y., Qiao, Y., Agrawala, M., Lin, D., Dai, B.: Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. In: ICLR (2024)
7. He, Z., Li, W., Zhang, T., Yuan, Y.: H 2 gm: A hierarchical hypergraph matching framework for brain landmark alignment. In: MICCAI (2023)
8. Hu, Q., Xiao, J., Chen, Y., Sun, S., Chen, J.N., Yuille, A., Zhou, Z.: Synthetic tumors make ai segment tumors better. NeurIPS Workshop (2022)
9. Li, C., Feng, B.Y., Fan, Z., Pan, P., Wang, Z.: Steganerf: Embedding invisible information within neural radiance fields. In: ICCV. pp. 441–453 (2023)
10. Li, C., Lin, M., Ding, Z., Lin, N., Zhuang, Y., Huang, Y., Ding, X., Cao, L.: Knowledge condensation distillation. In: ECCV (2022)

11. Li, C., Liu, H., Liu, Y., Feng, B.Y., Li, W., Liu, X., Chen, Z., Shao, J., Yuan, Y.: Endora: Video generation models as endoscopy simulators. ArXiv:2403.11050 (2024)
12. Li, C., Liu, X., Li, W., Wang, C., Liu, H., Yuan, Y.: U-kan makes strong backbone for medical image segmentation and generation. ArXiv:2406.02918 (2024)
13. Li, J., Xiong, C., Hoi, S.C.: Comatch: Semi-supervised learning with contrastive graph regularization. In: ICCV (2021)
14. Li, W., Chen, Z., Li, B., Zhang, D., Yuan, Y.: Htd: Heterogeneous task decoupling for two-stage object detection. TIP (2021)
15. Li, W., Guo, X., Yuan, Y.: Novel scenes & classes: Towards adaptive open-set object detection. In: ICCV. pp. 15780–15790 (2023)
16. Li, W., Liu, J., Han, B., Yuan, Y.: Adjustment and alignment for unbiased open set domain adaptation. In: CVPR. pp. 24110–24119 (2023)
17. Li, W., Liu, X., Yao, X., Yuan, Y.: Scan: Cross domain object detection with semantic conditioned adaptation. In: AAAI. pp. 1421–1428 (2022)
18. Li, W., Liu, X., Yuan, Y.: Scan++: Enhanced semantic conditioned adaptation for domain adaptive object detection. TMM (2022)
19. Li, W., Liu, X., Yuan, Y.: Sigma: Semantic-complete graph matching for domain adaptive object detection. In: CVPR (2022)
20. Li, W., Liu, X., Yuan, Y.: Sigma++: Improved semantic-complete graph matching for domain adaptive object detection. TPAMI (2023)
21. Liu, F., Tian, Y., Chen, Y., Liu, Y., Belagiannis, V., Carneiro, G.: Acpl: Anti-curriculum pseudo-labelling for semi-supervised medical image classification. In: CVPR. pp. 20697–20706 (2022)
22. Liu, X., Li, W., Yamaguchi, T., Geng, Z., Tanaka, T., Tsai, D.P., Chen, M.K.: Stereo vision meta-lens-assisted driving vision. ACS Photonics (2024)
23. Liu, X., Li, W., Yuan, Y.: Intervention & interaction federated abnormality detection with noisy clients. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 309–319. Springer (2022)
24. Liu, X., Li, W., Yuan, Y.: Decoupled unbiased teacher for source-free domain adaptive medical object detection. TNNLS (2023)
25. Liu, Y., Li, C., Yang, C., Yuan, Y.: Endogaussian: Gaussian splatting for deformable surgical scene reconstruction. ArXiv:2401.12561 (2024)
26. Liu, Y., Li, W., Liu, J., Chen, H., Yuan, Y.: Grab-net: Graph-based boundary-aware network for medical point cloud segmentation. TMI (2023)
27. Liu, Y., Liu, J., Yuan, Y.: Edge-oriented point-cloud transformer for 3d intracranial aneurysm segmentation. In: MICCAI (2022)
28. Özbey, M., Dalmaz, O., Dar, S.U., Bedel, H.A., Özturk, Ş., Güngör, A., Çukur, T.: Unsupervised medical image translation with adversarial diffusion models. TMI (2023)
29. Qiushi Yang, Xinyu Liu, Z.C., Yuan, Y.: Semi-supervised medical image classification with temporal knowledge-aware regularization. In: MICCAI (2022)
30. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al.: Imagenet large scale visual recognition challenge. IJCV (2015)
31. Sohn, K., Berthelot, D., Li, C.L., Zhang, Z., Carlini, N., Cubuk, E.D., Kurakin, A., Zhang, H., Raffel, C.: Fixmatch: Simplifying semi-supervised learning with consistency and confidence. In: NeurIPS (2020)
32. Sun, L., Wang, J., Huang, Y., Ding, X., Greenspan, H., Paisley, J.: An adversarial learning approach to medical image synthesis for lesion detection. JBHI **24**(8), 2303–2314 (2020)

33. Wang, Z., Liu, C., Zhang, S., Dou, Q.: Foundation model for endoscopy video analysis via large-scale self-supervised pre-train. In: MICCAI. pp. 101–111 (2023)
34. Xu, H., Zhang, Y., Sun, L., Li, C., Huang, Y., Ding, X.: Afsc: Adaptive fourier space compression for anomaly detection. ArXiv:2204.07963 (2022)
35. Xue, H., Sun, Y., Liu, B., Fu, J., Song, R., Li, H., Luo, J.: Clip-vip: Adapting pre-trained image-text model to video-language representation alignment. ArXiv:2209.06430 (2022)
36. Yang, Q., Li, W., Li, B., Yuan, Y.: Mrm: Masked relation modeling for medical image pre-training with genetics. In: ICCV (2023)
37. Zhang, Y., Li, C., Lin, X., Sun, L., Zhuang, Y., Huang, Y., Ding, X., Liu, X., Yu, Y.: Generator versus segmentor: Pseudo-healthy synthesis. In: MICCAI (2021)
38. Zheng, M., You, S., Huang, L., Luo, C., Wang, F., Qian, C., Xu, C.: Simmatchv2: Semi-supervised learning with graph consistency. In: ICCV. pp. 16432–16442 (2023)