TRUST: Efficient Abdominal Trauma Recognition via Image-to-Ultrasound-Video Transfer Learning

Enguang Wang[1,3,4⋆], Hao Zhou[2⋆], Shuo Gao[1,3,4], Tuo Liu[1,3,4], and Guangquan Zhou[1,3,4(✉)]

[1] School of Biological Science and Medical Engineering, Southeast University, Nanjing, China
guangquan.zhou@seu.edu.cn
[2] Nanjing Medical University First Affiliated Hospital, Nanjing, China
[3] Jiangsu Key Laboratory of Biomaterials and Devices, Southeast University, Nanjing, China
[4] State Key Laboratory of Digital Medical Engineering, Southeast University, Nanjing, China

Abstract. Abdominal ultrasound is indispensable for rapid, noninvasive trauma triage. However, interpreting the subtle dynamic cues embedded in continuous scanning is time-intensive and operator-dependent. Parameter-Efficient Image-to-Video Transfer Learning (PEIVTL), which efficiently adapts pre-trained image models to the video domain, notably through visual-textual alignment, offers a promising paradigm for ultrasound video analysis. Nevertheless, substantial spatiotemporal and semantic variations arising from physician-dependent scanning practices continue to limit the effectiveness and generalizability of this framework. We propose TRUST, a scan-aware PEIVTL framework that explicitly models fine-grained spatiotemporal variations to enable reliable ultrasound video understanding. First, we introduce a Cross-Frequency Collaborative Adapter (CFCA) that establishes mutual constraints between low- and high-frequency components, enhancing discriminative spatial feature extraction under heavy speckle corruption. Second, we design a Multi-Granularity Motion-Aware (MGMA) module that integrates local temporal convolutions with motion-prior-guided global self-attention, jointly capturing stable intra-view patterns and abrupt inter-view transitions to characterize complex scanning dynamics. Third, a Visual Query Semantic Aggregation (VQSA) module dynamically generates text prototypes conditioned on visual features, enabling adaptive visual-textual alignment robust to intra-class variability under diverse scanning conditions. Experiments on in-house ultrasound trauma datasets demonstrate that TRUST outperforms state-of-the-art methods by 9.63% with superior computational efficiency.

Keywords: Abdominal Ultrasound · Trauma Recognition · Image-to-Video Transfer Learning · Ultrasound Video Analysis.

⋆ Equal Contribution.

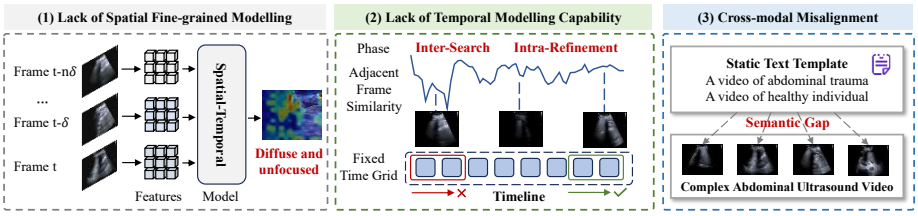

Fig. 1. Existing methods lack fine-grained spatial extraction, temporal dynamic modeling, and cross-modal matching capabilities in abdominal ultrasound applications.

1 Introduction

In emergency trauma care, abdominal ultrasound has emerged as the preferred imaging modality for rapid triage due to its non-invasive nature, portability, and real-time capabilities [12]. Nonetheless, accurate trauma assessment requires clinicians to interpret subtle, often transient, dynamic cues across continuous, multi-view scanning sequences, which is highly time-intensive and operator-dependent [18, 24]. These challenges underscore the need for automated, video-level analysis methods capable of modeling spatiotemporal information in ultrasound data. Although deep learning has achieved significant advances in medical video analysis [6, 20], its success critically depends on large-scale, temporally annotated datasets, which remain notably scarce in the ultrasound domain [19]. As an alternative, Parameter-Efficient Image-to-Video Transfer Learning (PEIVTL) [14, 17], which adapts pre-trained larger models to the video domain by leveraging visual–textual alignment in vision-language frameworks such as CLIP [11], offers a promising paradigm for ultrasound video analysis.

Despite the demonstrated effectiveness of PEIVTL [19, 16, 10], its direct application to abdominal ultrasound video analysis remains non-trivial due to the pronounced spatiotemporal and semantic variability introduced by operator-dependent scanning. First, speckle noise and imaging artifacts in ultrasound [1, 23] routinely obscure the subtle pathological cues that are essential for accurate diagnosis, demanding more discriminative and fine-grained spatial feature modeling. Moreover, the complex inter-scan variability arising from both rapid probe repositioning and gradual anatomical changes presents substantial challenges for characterizing the dual temporal dynamics intrinsic to ultrasound examinations. Most existing PEIVTL approaches rely on single-granularity temporal modeling, typically implemented via fixed-window local convolutions [7, 9, 25, 2], which are insufficient to simultaneously capture short-range motion continuity and longer-range cross-view transitions. Finally, considerable intra-class variation induced by heterogeneous scanning angles, probe pressures, and patient-specific factors limits the adaptability of static text templates [17, 13, 14] used for visual-textual alignment. Such rigidity can lead to suboptimal cross-modal feature alignment, thereby compromising classification robustness and generalization performance.

In this work, we propose TRUST, a scan-aware PEIVTL framework that progressively refines ultrasound video representations along three complementary axes—spatial detail, temporal dynamics, and cross-modal semantics—within a CLIP-based pipeline. Specifically, to enhance fine-grained spatial discrimination, we introduce the Cross–Frequency Collaborative Adapter (CFCA), which leverages frequency-domain interaction to model the spectral-informed complementarity between low- and high-frequency components. Building on these enriched spatial features, the Multi-Granularity Motion-Aware (MGMA) module aggregates temporal information through a dual-branch architecture: a local branch employing temporal convolutions for short-range intra-view consistency, and a global branch incorporating motion-prior position encoding to capture long-range inter-view scanning logic. The resulting visual representation then drives the Visual Query Semantic Aggregation (VQSA) module, which uses it as a query to dynamically synthesize instance-specific text prototypes from a diverse description repository via cross-modal attention, enabling adaptive visual–textual alignment robust to intra-class variability. The contributions of this paper are threefold: 1) To our best knowledge, this is the first work to adapt PEIVTL for abdominal ultrasound video analysis; 2) We propose TRUST, a unified framework that progressively integrates fine-grained spatial modeling, motion-aware temporal encoding, and visual-query-driven semantic alignment; 3) On an abdominal ultrasound trauma recognition dataset, TRUST achieves a 9.63% improvement over state-of-the-art methods with superior computational efficiency.

2 Methodology

As illustrated in Fig. 2, TRUST adopts a CLIP-style contrastive learning paradigm designed to progressively refine feature representations across spatial, temporal, and semantic dimensions. Specifically, within the visual branch, the CFCA mechanism is systematically integrated into each transformer layer to facilitate more discriminative and fine-grained feature extraction. In addition, the MGMA module is incorporated to enhance temporal dependency modeling and improve the representation of dynamic patterns. On the other hand, in the textual branch, the Text-Adapter and VQSA strategies are integrated to strengthen semantic representation and promote more precise cross-modal alignment.

2.1 Cross-Frequency Collaborative Adapter (CFCA)

Existing frequency-based approaches [23, 5] typically decompose features into high- and low-frequency components for independent processing. Although this strategy enhances frequency-specific modeling, it may amplify sensitivity to speckle noise inherent in ultrasound images, thereby limiting its effectiveness in abdominal ultrasound analysis. To this end, we propose CFCA, which explicitly models mutual constraints between low- and high-frequency components

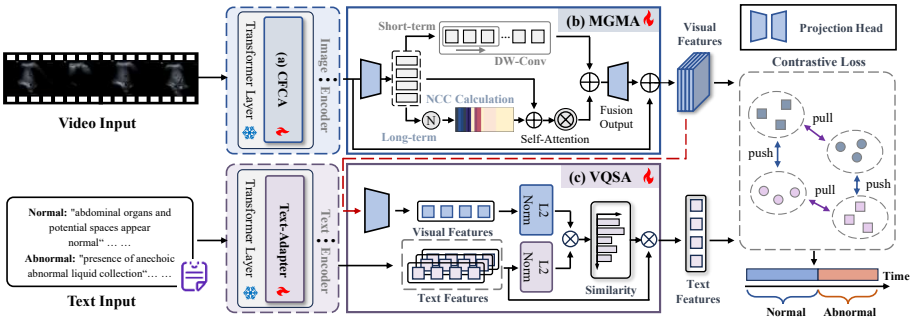

Fig. 2. Overview of the proposed TRUST. It includes: (a) CFCA, which extracts fine-grained spatial features; (b) MGMA, which models temporal dynamics in ultrasound videos; and (c) VQSA, which adaptively aggregates diverse textual descriptions.

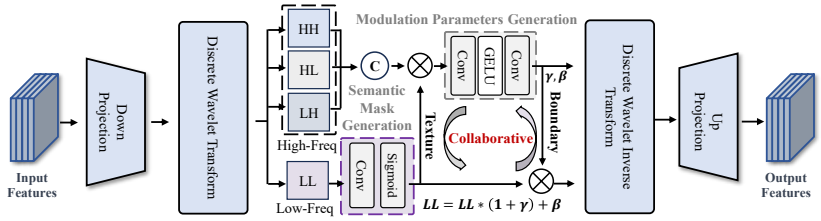

Fig. 3. Diagram of the CFCA module.

to promote cross-frequency information exchange, thereby extracting more discriminative features under challenging ultrasound scenarios [22].

As illustrated in Fig. 3, the input features are first decomposed via wavelet transform into a low-frequency sub-band (LL) encoding coarse structural patterns, and three high-frequency sub-bands (HL, LH, HH) capturing fine-grained detail variations. Since the low-frequency sub-band carries global semantic certainty, we use it to generate a spatial mask that gates the high-frequency components, suppressing noise-induced spurious responses while preserving diagnostically relevant details:

$$F_{High} = \sigma(Conv2D(F_{LL})) \odot Cat(F_{HL}, F_{LH}, F_{HH}), \tag{1}$$

Conversely, high-frequency features provide spatially precise cues that refine the low-frequency component's representational scope, preventing semantic overgeneralization. We thus derive adaptive scaling γ and offset β from them to sharpen low-frequency discrimination:

$$F_{Low} = F_{LL} \odot (1 + \gamma) + \beta, \tag{2}$$

Finally, the processed high-frequency and low-frequency features undergo inverse wavelet transformation and dimensionality projection to produce the final features $x_t = Proj(IWT(F_{Low}, F_{High}))$.

2.2 Multi-Granularity Motion-Aware (MGMA) Module

Ultrasound scanning exhibits dual temporal dynamics: smooth intra-view tissue motion and abrupt inter-view anatomical discontinuities from probe repositioning. Standard fixed-grid temporal layers [3, 4] apply uniform aggregation that inevitably conflates these two regimes, introducing noise at view boundaries. MGMA (Fig. 2(b)) disentangles them with a dual-branch architecture that enforces local continuity and calibrates global context via motion priors.

Short-range Branch (Intra-view Consistency). A depth-wise temporal convolution operates on the input sequence $\mathbf{X}$ as a temporal low-pass filter, yielding $\mathbf{X}_S = DWConv1D(\mathbf{X})$, which suppresses frame-level jitter while preserving subtle pathological motion continuity within stable views.

Long-range Branch (Inter-view Scanning Logic). To capture global context while mitigating the disruption caused by anatomical discontinuities, we introduce Motion-Prior Position Encoding (MPE). The core insight is that semantic displacement serves as a robust surrogate for physical probe movement. We define the relative physical coordinate p_t as the cumulative sum of semantic distances between adjacent frames within the causal sampling sequence. This mechanism dynamically stretches the temporal receptive field during rapid probe shifts and compresses it during stable scanning, aligning attention with the actual scanning logic:

$$p_t = \sum_{\tau=2}^{t} \left(1 - \frac{\mathbf{x}_\tau \cdot \mathbf{x}_{\tau-1}}{|\mathbf{x}_\tau||\mathbf{x}_{\tau-1}|}\right), \quad \mathbf{X}_L = \mathrm{MSA}\left(\{\mathbf{x}_t + \Phi(p_t)\}_{t=1}^{T}\right), \tag{3}$$

Finally, the output integrates both granularities via a residual projection $\mathbf{X}_{Out} = (\mathbf{X}_S + \mathbf{X}_L)\mathbf{W}_{up} + \mathbf{X}$.

2.3 Visual Query Semantic Aggregation (VQSA) Module

Given the visual embedding from MGMA, the final step is cross-modal alignment. However, static text templates [17, 13, 15] fail to capture the significant intra-class variations caused by diverse scanning angles and probe pressure. Since different scanning conditions produce visually distinct manifestations of the same pathology, the optimal textual representation should be conditioned on each instance's visual appearance. Based on this insight, VQSA uses the visual embedding as an *anchor* to dynamically aggregate the most relevant descriptions from a diverse textual repository (Fig. 2(c)). Formally, let $\mathbf{v} = \mathrm{Pool}(\mathbf{X}_{Out}) \in \mathbb{R}^d$ denote the final visual embedding. For the c-th category, we construct a fixed text bank using a structured slot-based prompting strategy, where each prompt is organized by category, modality, region, attribute, and temporal context. The generated candidates were manually verified according to FAST principles to exclude clinically ambiguous, redundant, or irrelevant descriptions before training. The resulting feature set $\mathcal{T}_c = \{\mathbf{t}_c^k\}_{k=1}^{K}$ is extracted via the text encoder equipped with a standard bottleneck adapter [14] for domain refinement. We

Table 1. Comparison of results within the abdominal ultrasound video dataset. Frames denotes the number of sampled frames × sampling interval. Modal indicates the input modality for the model: image alone is denoted as I, while simultaneous image and text input is denoted as $I + T$.

Frames	Method	Modal	Accuracy	Precision	Recall	F1-score	Jaccard
4×2	ST-Adapter[NIPS'22]	I	65.00	64.48	67.44	63.33	46.87
	AIM[ICLR'23]	I	68.77	61.80	60.57	60.98	46.31
	Surg-PEFT[TMI'25]	I	<u>70.76</u>	35.34	50.00	41.41	35.34
	Vita-CLIP[CVPR'23]	I+T	66.63	54.20	52.60	51.68	39.42
	M2-CLIP[AAAI'24]	I+T	69.69	<u>69.09</u>	<u>72.99</u>	<u>68.23</u>	<u>52.24</u>
	VLPA-CLIP[PR'25]	I+T	66.84	68.74	72.28	65.24	48.62
	TRUST[Ours]	I+T	**81.36**	**76.89**	**81.00**	**77.89**	**64.32**
8×2	ST-Adapter[NIPS'22]	I	66.74	59.28	58.84	59.02	44.18
	AIM[ICLR'23]	I	70.76	35.19	50.00	41.31	35.19
	Surg-PEFT[TMI'25]	I	65.26	56.22	55.15	55.23	41.29
	Vita-CLIP[CVPR'23]	I+T	71.06	63.44	51.80	46.30	37.89
	M2-CLIP[AAAI'24]	I+T	65.56	68.41	71.92	65.67	49.14
	VLPA-CLIP[PR'25]	I+T	<u>73.71</u>	<u>71.68</u>	<u>75.71</u>	<u>71.73</u>	<u>56.42</u>
	TRUST[Ours]	I+T	**83.34**	**78.63**	**82.71**	**79.78**	**66.85**

then employ cross-modal attention where $\mathbf{v}$ serves as the Query and $\mathbf{t}_c^k$ as Keys and Values, dynamically synthesizing an instance-specific category prototype:

$$\mathbf{p}_c(\mathbf{v}) = \sum_{k=1}^{K} \text{Softmax}\left(\frac{\mathbf{v}^\top \mathbf{t}_c^k}{\sqrt{d}}\right) \cdot \mathbf{t}_c^k, \tag{4}$$

Finally, following the CLIP paradigm, the classification logits are predicted by computing the cosine similarity between the visual embedding $\mathbf{v}$ and the synthesized dynamic prototype $\mathbf{p}_c(\mathbf{v})$.

2.4 Overall Loss

To enhance discriminative capabilities for downstream tasks while preserving pre-trained knowledge, we designed a hybrid objective function:

$$\mathcal{L}_{total} = \lambda \mathcal{L}_{con} + (1 - \lambda)\mathcal{L}_{cls}. \tag{5}$$

where λ is a balancing coefficient, which is set to 0.5, $\mathcal{L}_{con}$ adopts the standard symmetric InfoNCE formulation to achieve cross-modal alignment, whilst $\mathcal{L}_{cls}$ employs a standard cross-entropy loss based on image-text similarity logits, explicitly optimizing classification accuracy.

3 Experiments and Results

3.1 Datasets and Experimental Settings

For the abdominal ultrasound video trauma detection task, we utilized a proprietary dataset consisting of 294 videos. All frames in these videos were resized

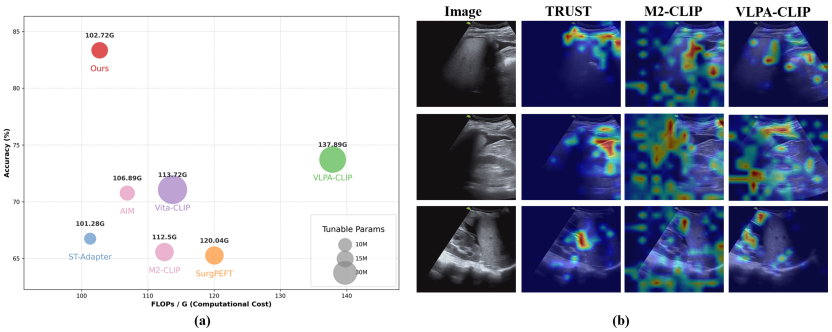

Fig. 4. (a) Comparison of computational complexity under an 8×2 sampling strategy, where the circle size represents the model's trainable parameter scale; (b) Heatmap of text-image similarity generated using the CLIP model.

to 224×224 resolution and annotated with labels indicating normal or abnormal stages. The videos were divided into training and test sets by patient, comprising 231 and 63 videos, respectively. For the number of texts corresponding to each category K, we set this value to 8 in the reported experimental results. For the evaluation metrics, we first selected the Jaccard coefficient (tIoU), which directly relates to interval localization. We then chose metrics for fine-grained recognition, including Accuracy, Recall, Precision, and the F1 score. For all experiments, both the image and text encoders used the weights of CLIP ViT-B/16, with the pre-trained encoders kept frozen during training. All experiments were conducted on 4 RTX 4090 GPUs, using a batch size of 32 and training for 36 epochs. We employ the AdamW optimizer and use a base learning rate of 1e-3.

3.2 Comparison with State-of-the-Art Methods.

We compared our approach with two categories of state-of-the-art PEIVTL methods: unimodal methods, including ST-Adapter [8], AIM [21], and SurgPEFT [19]; and multimodal methods, including Vita-CLIP [17], M2-CLIP [14], and VLPA-CLIP [13]. All methods were evaluated under identical conditions using two video sampling strategies: 4×2 and 8×2 (frames×interval). As shown in Table 1, under the 4×2 setting, TRUST achieved an accuracy of 81.36% and a Jaccard coefficient of 64.32%, surpassing the second-best method by 10.6 and 12.08 percentage points, respectively. This significant performance advantage underscores the necessity of our motion-aware design. Traditional methods suffer from feature confusion due to inherent anatomical variations in ultrasound, whereas TRUST effectively leverages motion prior information to model both short- and long-range temporal information.

Moreover, Fig. 4 (a) illustrates the complexity comparison under the 8×2 strategy, where TRUST achieves optimal classification performance while maintaining highly competitive computational efficiency. Fig. 4 (b) presents a qualitative analysis. The heatmap, based on image-text similarity, demonstrates that

Table 2. Experimental results of the ablation of each component of the model.

CFCA	Text-Adapter	MGMA	VQSA	Accuracy	Precision	Recall	F1-score	Jaccard
×	×	×	×	67.50	51.73	50.62	47.17	37.27
✓	×	×	×	70.96	69.17	73.01	68.66	52.78
✓	✓	×	×	75.38	71.77	72.93	72.27	57.68
✓	✓	✓	×	76.36	73.05	77.22	73.50	58.64
✓	✓	✓	✓	**83.34**	**78.63**	**82.71**	**79.78**	**66.85**

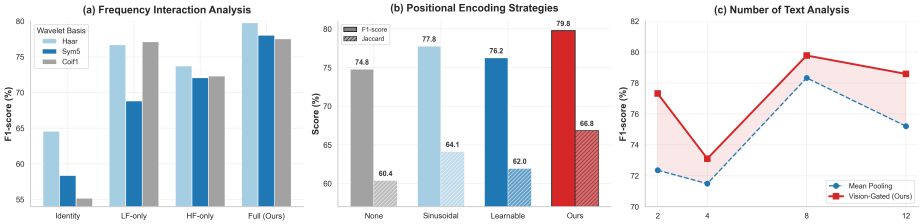

Fig. 5. Three analytical experiments conducted using an 8×2 sampling strategy: (a) analysis of frequency-domain interaction patterns and wavelet basis functions; (b) evaluation of the impact of different positional encoding schemes; and (c) investigation of the influence of varying text quantities and fusion strategies.

TRUST captures fine-grained spatial information, enabling precise localization of traumatic lesions. In contrast, the comparison method exhibits noticeable attention divergence or erroneous activation toward background noise, further validating the effectiveness of TRUST's cross-modal alignment.

3.3 Ablation Study.

Table 2 presents the results of ablation experiments for each component. The baseline model achieved an accuracy of only 67.50%. Incorporating CFCA increased accuracy to 70.96%, highlighting the crucial role of enhancing fine-grained spatial features. Adding Text-Adapter and MGMA on this foundation further improved model performance to 76.36%, demonstrating the effectiveness of text adaptation and motion perception-based temporal modeling. Most notably, the introduction of VQSA resulted in a significant performance boost, raising the accuracy and Jaccard score to 83.34% and 66.85%, respectively.

Furthermore, we conducted more detailed analytical experiments on each proposed module, with the results presented in Fig. 5. For CFCA, robustness was demonstrated across different wavelet basis functions, whereas interaction using sub-bands alone yielded suboptimal performance. Regarding positional encoding, our proposed motion-aware encoding achieved the best results. In terms of VQSA, our multi-text fusion strategy significantly outperformed mean pooling.

4 Conclusion

In this paper, we propose TRUST, a parameter-efficient image-to-video transfer learning framework specifically designed for abdominal ultrasound trauma detection. By integrating fine-grained spatial feature extraction, multi-granularity temporal information aggregation, and enhanced cross-modal alignment, our approach achieves highly efficient and accurate trauma detection. Notably, the method introduces a novel motion-aware positional encoding that effectively addresses semantic discontinuities caused by significant anatomical variations during ultrasound scanning, providing a new perspective for ultrasound video analysis tasks.

Acknowledgments. This work was supported in part by the National Natural Science Foundation of China (62371121), and in part by the Jiangsu Provincial Key Research and Development Program, China (BE2022827).

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

References

1. Deng, X., Wu, H., Zeng, R., Qin, J.: Memsam: Taming segment anything model for echocardiography video segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9622–9631 (2024)
2. Kong, Y., Fu, Y.: Human action recognition and prediction: A survey. International Journal of Computer Vision **130**(5), 1366–1401 (2022)
3. Lea, C., Flynn, M.D., Vidal, R., Reiter, A., Hager, G.D.: Temporal convolutional networks for action segmentation and detection. In: proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 156–165 (2017)
4. Lea, C., Vidal, R., Reiter, A., Hager, G.D.: Temporal convolutional networks: A unified approach to action segmentation. In: European conference on computer vision. pp. 47–54. Springer (2016)
5. Liu, J., Kong, L., Chen, G.: Improving sam for camouflaged object detection via dual stream adapters. arXiv preprint arXiv:2503.06042 (2025)
6. Liu, Y., Boels, M., Garcia-Peraza-Herrera, L.C., Vercauteren, T., Dasgupta, P., Granados, A., Ourselin, S.: Lovit: Long video transformer for surgical phase recognition. Medical Image Analysis **99**, 103366 (2025)
7. Ni, B., Peng, H., Chen, M., Zhang, S., Meng, G., Fu, J., Xiang, S., Ling, H.: Expanding language-image pretrained models for general video recognition. In: European conference on computer vision. pp. 1–18. Springer (2022)
8. Pan, J., Lin, Z., Zhu, X., Shao, J., Li, H.: St-adapter: Parameter-efficient image-to-video transfer learning. Advances in Neural Information Processing Systems **35**, 26462–26477 (2022)
9. Park, J., Lee, J., Sohn, K.: Dual-path adaptation from image to video transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2203–2213 (2023)

10. Pei, W., Tan, Q., Lu, G., Tian, J., Yu, J.: D2st-adapter: Disentangled-and-deformable spatio-temporal adapter for few-shot action recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 11317–11326 (2025)
11. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PmLR (2021)
12. van Randen, A., Laméris, W., van Es, H.W., van Heesewijk, H.P., van Ramshorst, B., Ten Hove, W., Bouma, W.H., van Leeuwen, M.S., van Keulen, E.M., Bossuyt, P.M., et al.: A comparison of the accuracy of ultrasound and computed tomography in common diagnoses causing acute abdominal pain. European radiology **21**(7), 1535–1545 (2011)
13. Wang, H., Liu, F., Jiao, L., Wang, J., Li, S., Li, L., Chen, P., Liu, X., Ma, W.: Vlpa-clip: Video language prompting and adapting clip for efficient video action recognition. Pattern Recognition p. 111770 (2025)
14. Wang, M., Xing, J., Jiang, B., Chen, J., Mei, J., Zuo, X., Dai, G., Wang, J., Liu, Y.: A multimodal, multi-task adapting framework for video action recognition. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 5517–5525 (2024)
15. Wang, M., Xing, J., Liu, Y.: Actionclip: A new paradigm for video action recognition. arXiv preprint arXiv:2109.08472 (2021)
16. Wang, Y., Liu, M., Song, X., Nie, L.: Tr-adapter: Parameter-efficient transfer learning for video question answering. IEEE Transactions on Multimedia (2024)
17. Wasim, S.T., Naseer, M., Khan, S., Khan, F.S., Shah, M.: Vita-clip: Video and text adaptive clip via multimodal prompting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 23034–23044 (2023)
18. Xu, J., Li, X., Yue, C., Wang, Y., Guo, Y.: Sam-mpa: Applying sam to few-shot medical image segmentation using mask propagation and auto-prompting. arXiv preprint arXiv:2411.17363 (2024)
19. Yang, S., Cai, Z., Luo, L., Ma, N., Xu, S., Chen, H.: Surgpetl: Parameter-efficient image-to-surgical-video transfer learning for surgical phase recognition. IEEE Transactions on Medical Imaging (2025)
20. Yang, S., Luo, L., Wang, Q., Chen, H.: Surgformer: Surgical transformer with hierarchical temporal attention for surgical phase recognition. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 606–616. Springer (2024)
21. Yang, T., Zhu, Y., Xie, Y., Zhang, A., Chen, C., Li, M.: Aim: Adapting image models for efficient video action recognition. arXiv preprint arXiv:2302.03024 (2023)
22. Ye, S., Chen, X., Zhang, Y., Lin, X., Cao, L.: Escnet: Edge-semantic collaborative network for camouflaged object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 20053–20063 (2025)
23. Zhang, W., Wu, H., Qin, J.: Domesticating sam for breast ultrasound image segmentation via spatial-frequency fusion and uncertainty correction. In: European Conference on Computer Vision. pp. 20–37. Springer (2024)
24. Zhou, N., Zou, K., Ren, K., Luo, M., He, L., Wang, M., Chen, Y., Zhang, Y., Chen, H., Fu, H.: Medsam-u: Uncertainty-guided auto multi-prompt adaptation for reliable medsam. IEEE Transactions on Circuits and Systems for Video Technology (2025)

25. Zhu, Y., Li, X., Liu, C., Zolfaghari, M., Xiong, Y., Wu, C., Zhang, Z., Tighe, J., Manmatha, R., Li, M.: A comprehensive study of deep video action recognition. arXiv preprint arXiv:2012.06567 (2020)