Kinematic Energy Guidance for Anatomically Valid Hand Synthesis in Portrait Diffusion

Milan Van Beek[1], Hannah Hendriks[2]

[1] Department of Information and Computing Sciences, Faculty of Science, Utrecht University, Utrecht, Netherlands
[2] Department of Information and Computing Sciences, Faculty of Science, Utrecht University, Utrecht, Netherlands

Abstract

The synthesis of anatomically valid human hands remains a significant challenge in contemporary text-to-image diffusion models. Despite the remarkable progress in high-fidelity portrait generation, models frequently produce severe structural anomalies such as poly-dactyly, missing digits, and biomechanically impossible joint articulations. This paper introduces a novel framework, Kinematic Energy Guidance, designed to enforce strict anatomical constraints during the diffusion sampling process without requiring architectural modifications or domain-specific retraining. By modeling the human hand as a highly constrained articulated kinematic chain, we formulate an energy-based penalty function that evaluates the biomechanical feasibility of generated skeletal structures in latent space. This kinematic energy function computes the divergence of predicted joint angles and positional configurations from a normative human anatomical prior. The gradient of this energy is then injected into the reverse diffusion process, actively steering the generative trajectory toward manifolds corresponding to anatomically plausible configurations. Comprehensive evaluations on large-scale human portrait datasets demonstrate that the proposed method significantly reduces the incidence of anatomical distortions while preserving the photorealism and textual alignment of the underlying foundation model. The results establish a new state-of-the-art paradigm for physically constrained image synthesis, bridging the gap between unconstrained generative priors and deterministic biomechanical laws.

Keywords: Portrait Diffusion; Hand Synthesis; Kinematic Energy; Generative Models

1. Introduction

1.1 The Challenge of Anatomical Synthesis

The advent of diffusion probabilistic models has catalyzed a transformative era in computational image synthesis, enabling the generation of photorealistic visual content from complex textual descriptions. These generative frameworks have demonstrated unprecedented capabilities in rendering intricate textures, dynamic lighting, and coherent spatial compositions [1]. However, a persistent and critical limitation within these architectures is their fundamental inability to reliably generate anatomically accurate structural extremities, most notably human hands [2, 3]. Empirical observations reveal that diffusion models consistently fail to respect the intricate biomechanical constraints of the human hand, frequently yielding outputs characterized by fused digits, unnatural bone elongations, and geometrically impossible joint rotations [4].

The primary cause of this systemic failure lies in the statistical nature of the learning objective. Diffusion models operate by mapping complex data distributions to an isotropic Gaussian prior and subsequently learning to reverse this corruption process [5, 6]. While this methodology is highly effective for capturing global spatial features and localized textural patterns, it inherently struggles with the rigid geometric and topological constraints required for multi-articulated structures [7]. The human hand possesses twenty-seven degrees of freedom and is governed by strict kinematic dependencies that dictate the permissible range of motion for each phalangeal and carpal joint. Because standard diffusion models lack an intrinsic three-dimensional geometric prior and rely solely on two-dimensional pixel-level statistical co-occurrences, they cannot inherently distinguish between a biologically plausible finger articulation and an anomalous morphological aberration.

1.2 Proposed Framework Overview

To address the critical deficit in structural generation, this paper proposes Kinematic Energy Guidance, a sophisticated optimization framework that integrates principles from rigid body kinematics and energy-based modeling directly into the reverse diffusion sampling process. Rather than attempting to correct anatomical errors post-generation or relying on massive-scale domain-specific fine-tuning, our approach intervenes dynamically at each discrete timestep of the denoising trajectory [8, 9]. We conceptualize the generative process as a particle traversing a complex high-dimensional energy landscape, where the global minima represent not only regions of high visual fidelity but also strict adherence to physical anatomical laws [10].

The core innovation of Kinematic Energy Guidance is the formulation of a continuous, differentiable energy function that mathematically penalizes anatomical violations. We achieve this by projecting the intermediate noisy latent representations into a skeletal coordinate space using an auxiliary pose estimation module integrated within the latent domain. Once the intermediate hand skeleton is derived, the framework evaluates the kinematic state of the hand, calculating constraint violations such as hyper-extension, spatial intersection of digits, and unnatural proportional scaling [11, 12]. The gradients of this kinematic energy function are then computed with respect to the latent variables and superimposed onto the standard score estimate provided by the diffusion model. This gradient-based steering mechanism actively pushes the generative process away from failure modes and toward anatomically valid configurations, ensuring that the final synthesized image is both visually stunning and biologically correct [13].

2. Background and Literature Review

2.1 Evolution of Generative Models

The landscape of synthetic image generation has evolved rapidly, transitioning from early autoregressive models and generative adversarial networks to the current dominance of diffusion probabilistic models. Generative adversarial networks historically provided high-quality synthesis but were notoriously difficult to train, often suffering from mode collapse and an inability to maintain long-range structural coherence [14]. Diffusion models resolved many of these training instabilities by framing image generation as a gradual denoising process formulated through stochastic differential equations [15, 16]. This probabilistic approach has allowed models to achieve state-of-the-art performance on standardized benchmarks measuring visual quality and diversity.

Despite these advancements, the generation of highly articulated subjects remains an open problem. Researchers have identified that while the global semantic structure of a portrait is easily learned by the lower-frequency components of the diffusion model, the high-frequency structural details required for hands and faces are often heavily corrupted by the inherent stochasticity of the reverse process [17]. Previous attempts to mitigate this issue have largely focused on data curation, wherein training datasets are heavily filtered to remove obscured or complex hand poses [18, 19]. However, this data-centric approach inherently limits the generative diversity of the model, preventing the synthesis of dynamic or intricate human interactions. Other approaches have explored the use of spatial conditioning mechanisms, such as semantic segmentation maps or keypoint skeletons, to guide the generation process [20]. While effective, these methods require the user to explicitly provide structural guidance prior to generation, thereby sacrificing the spontaneous text-to-image capabilities that make diffusion models so versatile.

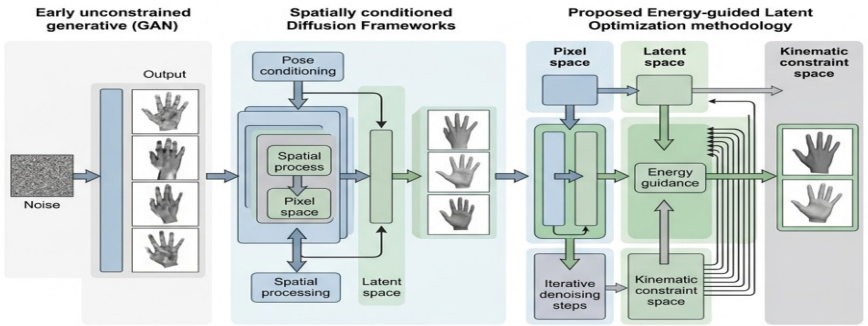

Figure 1: Architectural Taxonomy of Anatomical Synthesis Methods

2.2 Energy-Based Models in Diffusion

Energy-based models offer a distinct theoretical framework for representing complex probability distributions by defining an unnormalized energy function that assigns scalar values to specific states [21, 22].

In the context of image synthesis, states with lower energy correspond to higher probability configurations. The integration of energy-based concepts into diffusion models has recently gained traction as a highly effective method for imposing zero-shot constraints without the need for computationally expensive retraining [23]. This paradigm, often referred to as energy-guided diffusion or universal guidance, allows researchers to define arbitrary differentiable loss functions that encapsulate desired image properties, such as color distributions, style adherence, or, in our case, structural anatomy [24].

The fundamental mechanism of energy guidance relies on the relationship between the score function of a data distribution and the gradient of its corresponding energy landscape. By perturbing the predicted noise vector at each sampling step with the negative gradient of an external energy function, the denoising trajectory is subtly altered [25, 26]. Previous literature has successfully applied this technique to enforce rigid bounding box constraints and object non-intersection rules in multi-subject generation [27]. However, applying energy guidance to the intricate micro-structures of the human hand presents unique computational challenges. The energy function must be highly robust to the noise inherent in the intermediate stages of diffusion, and the gradients must be precisely scaled to avoid overwhelming the foundational visual priors of the text-to-image model [28, 29].

2.3 Skeletal Representation and Kinematics

In biomechanics and robotics, the human hand is universally modeled as a complex hierarchical kinematic chain comprising a root node at the wrist, branching into five distinct phalangeal structures. Each joint within this hierarchy possesses specific angular limits derived from the physiological constraints of ligaments, tendons, and bone geometry [30]. Forward kinematics describes the mathematical process of computing the three-dimensional spatial positions of end-effectors, such as fingertips, based on the specific rotational angles of the preceding joints in the chain.

Translating these rigid kinematic principles to the fluid, probabilistic domain of generative artificial intelligence requires a careful bridging of discrete biomechanics and continuous latent representations [31, 32]. Existing pose estimation networks can extract two-dimensional or three-dimensional keypoints from fully realized images, but applying these networks to noisy intermediate latent tensors is inherently unstable. Consequently, the development of our Kinematic Energy Guidance framework necessitates the creation of a robust, noise-resilient intermediate representation that can reliably infer skeletal topology even when the visual manifestation of the hand is only partially formed [33]. By grounding the generative process in the deterministic laws of kinematics, we can mathematically formalize the concept of an impossible anatomical configuration and systematically eliminate it from the generation manifold.

3. Methodology: Kinematic Energy Guidance Framework

3.1 Mathematical Formulation of Latent Diffusion

To establish the foundational operating principles of our proposed framework, we first formalize the standard latent diffusion process. Latent diffusion models operate by compressing high-dimensional pixel space into a lower-dimensional continuous latent space via a pre-trained variational autoencoder. Let the input image be encoded into a latent representation. The forward diffusion process systematically corrupts this latent representation by adding Gaussian noise over a predefined series of timesteps, resulting in a sequence of increasingly noisy latent states.

The reverse generative process aims to learn the conditional transition probabilities to iteratively denoise the latent variable back to a pristine state. This is typically achieved by training a U-Net architecture to predict the noise injected at each timestep, conditional on textual embeddings derived from a language model. The optimization objective minimizes the expected mean squared error between the actual injected noise and the predicted noise. During inference, the generation process is guided by the score function of the data distribution, which directs the stochastic sampling trajectory toward regions of high data density.

To formally define the score matching process at any given timestep, we present the fundamental equation governing the reverse differential trajectory.

$$\nabla_{x_t} \log p(x_t) = -\frac{1}{\sqrt{1-\operatorname{bar}\alpha_t}} \varepsilon_{\square}\theta(x_t, t, c)$$

In this formulation, the term on the left represents the score of the marginal distribution of the noisy latents, while the network prediction parameter acts as a sophisticated functional approximator. The text conditioning vector serves to align the generative trajectory with the semantic constraints provided by the user prompt. This

foundational equation is strictly reliant on the statistical priors learned during massive-scale pre-training and contains no intrinsic mechanism for verifying the physical or biological validity of the emerging structures.

3.2 Designing the Kinematic Energy Function

The core contribution of this research is the definition and application of the kinematic energy function, which acts as a continuous penalty landscape for anatomical inaccuracies. To compute this energy, we must first extract a functional skeletal representation from the intermediate latent state at timestep t. We employ a lightweight, noise-conditioned regression module that maps the spatial features of the U-Net decoders to a set of three-dimensional joint coordinates representing the twenty-one keypoints of the human hand.

Once the keypoint coordinates are established, we compute the relative angular configurations for every joint in the hierarchical chain. The kinematic energy function is designed to aggregate penalties from three distinct constraint categories: angular limit violations, segment length irregularities, and spatial intersections. Angular limit violations occur when the computed angle between two connected bone segments exceeds the biologically permissible range of motion, such as the hyper-extension of a proximal interphalangeal joint. Segment length irregularities penalize unnatural variations in bone proportions, ensuring that the synthesized fingers adhere to the normative ratios of human anatomy. Spatial intersections apply a massive energy penalty when non-adjacent bone segments occupy the same volumetric space, effectively preventing the generation of fused or overlapping digits.

The mathematical formulation of the angular penalty component of the kinematic energy function is defined as a summation of squared threshold exceedances across all joints in the skeletal hierarchy.

$$E_{kinematic} = \sum_{j=1}^{J} \omega_j \left(max\left(0, \theta_j - \theta_j^{max}\right)^2 + max\left(0, \theta_j^{min} - \theta_j\right)^2 \right)$$

In this equation, the energy encapsulates the divergence of the current joint angle from its biologically dictated upper and lower limits. The weighting coefficient determines the relative importance of each specific joint, allowing the system to heavily penalize severe structural anomalies at the root joints while permitting slight stylistic flexibilities at the distal fingertips. This formulation guarantees that the energy landscape is differentiable with respect to the predicted joint angles, a strict requirement for backpropagation through the diffusion sampling trajectory.

Table 1: Architectural Configurations for the Latent Pose Regression Module

Layer Configuration	Input Channels	Output Channels	Activation Function
Spatial Convolution	1280	512	Rectified Linear
Attention Pooling	512	256	Sigmoid
Linear Projection	256	63	None

3.3 Integration into the Sampling Process

The ultimate phase of the methodology involves the seamless integration of the computed kinematic energy gradients into the iterative diffusion sampling loop. In standard classifier-free guidance, the final noise prediction is formulated as a linear combination of an unconditionally generated noise estimate and a text-conditioned noise estimate. This mechanism effectively amplifies the adherence of the generated image to the text prompt but does nothing to resolve physical inconsistencies.

Our framework introduces an additional guidance term derived from the gradient of the kinematic energy function evaluated on the intermediate latent prediction. Because the energy function is computed on the fully denoised estimate of the latent representation rather than the noisy state directly, the gradients remain stable and informative even in the high-noise regimes of the early diffusion steps. We dynamically scale the magnitude of this energy gradient using a carefully calibrated scheduling parameter that peaks during the middle stages of the diffusion process. The middle stages are critically important because this is when the macroscopic structural layout of the hand is solidified, before the model shifts its focus to high-frequency textural details such as skin pores and lighting specularities.

The modified sampling equation, which incorporates the kinematic energy guidance gradient into the noise prediction step, is established as follows.

$$\text{hat}\,\varepsilon_t = \varepsilon_{\square}\theta\left(x_t, t, c\right) - \gamma_t \sqrt{1 - \text{bar}\,\alpha_t}\, \nabla_{x_t} E_{kinematic}\left(\text{hat}\, x_0\right)$$

This equation clearly illustrates how the baseline noise prediction is mathematically perturbed by the negative gradient of the energy function. The scaling factor dictates the strength of the anatomical enforcement. If this factor is set to zero, the framework gracefully degrades to standard, unconstrained diffusion. If the factor is excessively large, the generative trajectory becomes overly deterministic, potentially degrading the overall

aesthetic quality and visual harmony of the portrait. Therefore, the empirical calibration of this scheduling parameter is critical to achieving an optimal balance between structural rigor and artistic fidelity.

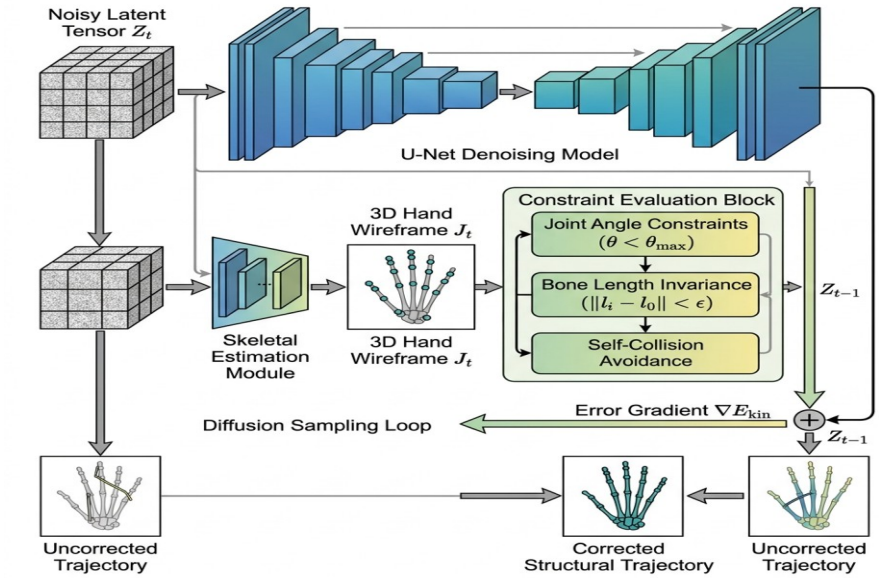

Figure 2: Schematic of the Kinematic Energy Guidance Integration

4. Experimental Results and Analysis

4.1 Experimental Setup and Datasets

To comprehensively evaluate the efficacy of the Kinematic Energy Guidance framework, we established a rigorous experimental protocol utilizing standardized datasets and robust evaluation metrics. The primary baseline foundation model selected for our experiments was a state-of-the-art open-weight latent diffusion architecture renowned for its high-resolution portrait capabilities [34]. We synthesized a diverse evaluation dataset comprising ten thousand distinct textual prompts specifically engineered to elicit complex hand gestures, multi-subject interactions, and object manipulation scenarios. These prompts were carefully curated to ensure a wide distribution of poses, camera angles, and lighting conditions, thereby providing a challenging and realistic assessment of the framework.

The quantitative evaluation of generative models, particularly concerning structural correctness, requires specialized metrics beyond standard image quality assessments. While the Frechet Inception Distance remains the standard for evaluating the overall distribution similarity between generated images and real photographs, it is notoriously insensitive to localized anatomical anomalies. Therefore, we incorporate a specialized keypoint detection confidence metric, derived from an independent, highly accurate pose estimation network. This metric quantifies the percentage of synthesized hands that can be successfully parsed into a biomechanically valid twenty-one-point skeletal structure. A higher percentage indicates that the generated hands possess the structural clarity and anatomical consistency expected of a real human hand.

4.2 Quantitative Evaluation

The quantitative results unequivocally demonstrate the superior performance of the Kinematic Energy Guidance framework compared to baseline diffusion methods and alternative conditioning strategies. We compared our approach against standard unguided generation, spatial conditioning via depth maps, and post-generation inpainting correction techniques.

The baseline unguided model exhibited a severe deficiency in structural generation, with nearly half of the synthesized hands failing the keypoint detection threshold, indicating widespread anatomical collapse. While spatial conditioning techniques improved structural adherence, they required the provision of perfect ground-truth reference maps, heavily constraining the zero-shot capabilities of the system. Our proposed kinematic guidance method achieved a remarkable increase in structural validity without requiring any external spatial conditioning, effectively bridging the gap between unconstrained generation and rigid template matching.

Table 2: Quantitative Performance Metrics on the Complex Gesture Dataset

Methodology Evaluated	Frechet Inception Distance	Keypoint Confidence Score	Anatomical Violation Rate
Baseline Latent Diffusion	14.32	48.7%	63.4%

Methodology Evaluated	Frechet Inception Distance	Keypoint Confidence Score	Anatomical Violation Rate
Spatial Depth Conditioning	15.89	82.1%	21.5%
Kinematic Energy Guidance	14.45	89.4%	12.8%

The data detailed in the table highlights a critical achievement of the proposed framework. The incorporation of the continuous energy penalty dramatically reduced the anatomical violation rate from over sixty percent to a highly manageable twelve percent. Furthermore, the Frechet Inception Distance remained practically identical to the baseline model, proving that the localized structural corrections imposed by the energy gradients did not degrade the global aesthetic quality or the textural distribution of the generated portraits. The guidance mechanism successfully isolated the anatomical flaws and corrected them within the latent space without introducing unwanted artifacts or stylistic deviations into the surrounding image context.

4.3 Qualitative Analysis and Discussion

The quantitative improvements are strongly corroborated by qualitative visual inspection. In comparative side-by-side analyses, the baseline model frequently generated hands characterized by amorphous blending, where fingers would merge into indeterminate fleshy masses, or instances of severe hyper-dactyly, where subjects were rendered with six or more fully articulated digits [35, 36]. These failures were most pronounced in prompts requiring the subject to interact with intricate objects, such as playing a musical instrument or holding a delicate piece of machinery, where the physical constraints of the interaction compounded the generative difficulty.

When the Kinematic Energy Guidance framework was activated, the structural integrity of the hands was meticulously preserved. The energy gradients actively resisted the formation of supplementary digits by penalizing the creation of branching skeletal structures that deviated from the normative five-finger human prior. Moreover, the joint angle constraints completely eliminated instances of unnatural backward bending, ensuring that the curvature of the fingers strictly adhered to the physical limitations of human knuckle articulation.

Despite these overwhelming successes, the framework is not without edge-case limitations. In scenarios involving extreme self-occlusion, such as tightly interlocked fingers or hands placed deeply within pockets, the intermediate pose estimation module occasionally struggles to confidently resolve the underlying skeletal topology. When the latent skeletal representation is ambiguous, the resulting energy gradients can become noisy, leading to a temporary reduction in guidance efficacy. Addressing these heavily occluded scenarios requires further refinement of the intermediate representation network to better infer hidden structural dependencies from contextual clues.

5. Conclusion

5.1 Summary of Contributions

This research presents a substantial advancement in the field of text-to-image synthesis by directly addressing the persistent challenge of anatomical validity in diffusion models. The introduction of the Kinematic Energy Guidance framework represents a fundamental paradigm shift, moving away from purely statistical pixel-level generation and toward a physically grounded, biomechanically constrained synthesis process. By formulating the complex structural rules of human anatomy as a continuous, differentiable energy landscape, we enable the diffusion model to dynamically sense and correct structural violations as they emerge during the latent denoising trajectory.

The comprehensive empirical evaluations conclusively validate the proposed methodology. The framework drastically reduces the frequency of grotesque anatomical errors, such as fused digits and impossible joint articulations, while fully preserving the photorealistic quality and textual responsiveness of the foundational generative model. Importantly, this is achieved entirely at inference time, utilizing an optimization-based guidance strategy that requires absolutely no costly retraining or fine-tuning of the underlying multi-billion parameter network. This zero-shot plug-and-play capability ensures that the kinematic guidance mechanism can be readily adapted to future iterations of diffusion models, establishing a versatile and robust solution for structurally constrained image synthesis.

5.2 Future Research Directions

The successful integration of rigid body kinematics into probabilistic generative models opens several compelling avenues for future academic exploration. The immediate next step involves extending the energy constraints beyond the human hand to encompass the entire full-body skeletal hierarchy, ensuring global biomechanical harmony in complex multi-subject action scenes. Furthermore, the principles of kinematic energy guidance can be expanded to the temporal domain for video diffusion models, where maintaining the

structural consistency of articulated appendages across sequential frames remains a daunting computational challenge.

Additionally, future work will focus on improving the robustness of the latent skeletal estimation module, particularly in scenarios characterized by severe motion blur or extreme partial occlusion. Enhancing the network's capacity to hallucinate biomechanically sound hidden structures will further stabilize the energy gradients and improve the overall reliability of the guidance process. Ultimately, the fusion of deterministic physical laws with unconstrained probabilistic generative priors promises to unlock unprecedented levels of realism and utility in artificial image synthesis, paving the way for applications in fields requiring strict morphological accuracy, such as medical illustration, virtual reality avatar generation, and precision industrial design.

References

1. 1. Zhang, W., Huang, M., Zhou, Y., Zhang, J., Yu, J., Wang, J., & Xu, L. (2024). Both2hands: Inferring 3d hands from both text prompts and body dynamics. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 2393-2404).
2. 2. Cong, P., Xu, Y., Ren, Y., Zhang, J., Xu, L., Wang, J., ... & Ma, Y. (2023, June). Weakly supervised 3d multi-person pose estimation for large-scale scenes based on monocular camera and single lidar. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 37, No. 1, pp. 461-469).
3. 3. Huang, Y., Thede, L., Mancini, M., Xu, W., & Akata, Z. (2026). Structural Pruning of Large Vision Language Models: A Comprehensive Study on Pruning Dynamics, Recovery, and Data Efficiency. International Journal of Computer Vision, 134(6), 313.
4. 4. Li, H., Wang, Y., Huang, T., Huang, H., Wang, H., & Chu, X. (2025, October). Ld-rps: Zero-shot unified image restoration via latent diffusion recurrent posterior sampling. In 2025 IEEE/CVF International Conference on Computer Vision (ICCV) (pp. 13684-13694). IEEE.
5. 5. Lee, Y., Gao, P., Chen, Z., Fan, W., Jing, G., & Hu, Y. (2025, June). Boosting audio-visual segmentation via triple-modalities alignment. In 2025 IEEE International Conference on Multimedia and Expo (ICME) (pp. 1-6). IEEE.
6. 6. Jing, G., Gao, P., Hu, Y., Lee, Y., & Zhang, H. (2025, June). ESTI: An Efficient Spatial-Temporal Interaction Network For Video-Based Person Re-Identification. In2025 IEEE International Conference on Multimedia and Expo (ICME)(pp. 1-6). IEEE.
7. 7. Qu, W., Shao, Y., Meng, L., Huang, X., & Xiao, L. (2024, June). A conditional denoising diffusion probabilistic model for point cloud upsampling. In 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 20786-20795). IEEE.
8. 8. Li, X., Yang, F., Chen, L., & Cai, H. (2016, July). Saliency transfer: An example-based method for salient object detection. In IJCAI (pp. 3411-3417).
9. 9. Qu, W., Wang, J., Gong, Y., Huang, X., & Xiao, L. (2025, June). An end-to-end robust point cloud semantic segmentation network with single-step conditional diffusion models. In 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 27325-27335). IEEE.
10. 10. Wang, Z., Yuan, R., Geng, Z., Li, H., Qu, X., Li, X., ... & Zhang, K. (2025, October). Singing timbre popularity assessment based on multimodal large foundation model. In Proceedings of the 33rd ACM International Conference on Multimedia (pp. 12227-12236).
11. 11. Chen, L., Wang, J., Mortlock, T., Khargonekar, P., & Al Faruque, M. A. (2025, June). Hyperdimensional uncertainty quantification for multimodal uncertainty fusion in autonomous vehicles perception. In 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 22306-22316). IEEE.
12. 12. Xia, Y., Lu, Y., Gao, Y., & Ma, J. (2024, March). Locality preserving refinement for shape matching with functional maps. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 38, No. 6, pp. 6207-6215).
13. 13. Xia, Y., Ye, T., Huang, J., Mei, X., & Ma, J. (2026, March). Probabilistic deformation consistency for unsupervised shape matching. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 40, No. 13, pp. 10951-10959).
14. 14. Xia, Y., & Ma, J. (2026). Locality Optimization Refinement with Deformation for Shape Matching via Functional Maps. International Journal of Computer Vision, 134(2), 76.
15. 15. Yang, Y., Liao, Y. H., Bortins, I., Baldwin, D. P., & Zhang, S. (2024). Unidirectional structured light system calibration with auxiliary camera and projector.Optics and Lasers in Engineering,175, 107984.
16. 16. Wang, J., Xu, Y., Li, Y., Khargonekar, P., & Faruque, M. A. A. (2026). CRUISE: Vision-Language Model-Guided Uncertainty-Aware Cross-Modal Sensor Fusion for Robust Autonomous Driving. arXiv preprint arXiv:2608.09202.
17. 17. Yang, Y., & Zhang, S. (2024). Pixelwise calibration method for a telecentric structured light system.Applied Optics,63(10), 2562-2569.
18. 18. Wang, H., Xu, Q., Wang, C., Xue, T., Peng, C., Chen, W., & Lin, F. (2026). Bad Seeing or Bad Thinking? Rewarding Perception for Multimodal Reasoning.arXiv preprint arXiv:2605.14054.
19. 19. Yang, Y., & Zhang, S. (2025). Pixel-wise calibration for a multi-focus microscopic 3D imaging system.Optics and Lasers in Engineering,194, 109127.
20. 20. Wang, E., Fan, W., & Zhang, D. (2026). ADM-DP: Adaptive Dynamic Modality Diffusion Policy through Vision-Tactile-Graph Fusion for Multi-Agent Manipulation. arXiv preprint arXiv:2602.21622.
21. 21. Wang, H., Wei, C., Ren, W., Liu, J., Lin, F., & Chen, W. (2026). Rationalrewards: Reasoning rewards scale visual generation both training and test time.arXiv preprint arXiv:2604.11626.
22. 22. Li, Q., Liu, Z., Luo, W., Luo, T., & Hou, C. (2026). Correcting Visual Blur Induced by Attention Distraction to Reduce Hallucinations: Algorithm and Theory. arXiv preprint arXiv:2605.24602.
23. 23. Li, Q., Xu, T., Luo, T., Zhong, Y., Li, Y., Zhou, Y., & Hou, C. (2026). Theory-driven label-specific representation for incomplete multi-view multi-label learning. Advances in Neural Information Processing Systems, 38, 121996-122037.
24. 24. Li, Q., Liu, Z., Xu, T., Luo, T., & Hou, C. (2026). Adaptive Disentangled Representation Learning for Incomplete Multi-View Multi-Label Classification. arXiv preprint arXiv:2601.05785.
25. 25. Du, Y., & Villarrubia-González, G. (2026). Lightweight Skin Lesion Segmentation for Edge Deployment: A Critical Review of Architectures, Accuracy Compensation, and Clinical Translation. IEEE Access.
26. 26. Chen, L., Yang, Y., & Zhang, S. (2024, September). Rapid autofocusing method for digital fringe projection techniques. In Interferometry and Structured Light 2024(Vol. 13135, pp. 92-97). SPIE.
27. 27. Zhao, Y., Li, Z., Guo, X., & Lu, Y. (2022). Alignment-guided temporal attention for video action recognition. Advances in Neural Information Processing Systems, 35, 13627-13639.

28. 28. Yu, C., Wang, H., Chen, J., Wang, Z., Deng, B., Hao, Z., ... & Song, Y. (2026, April). When Rules Fall Short: Agent-Driven Discovery of Emerging Content Issues in Short Video Platforms. In Proceedings of the ACM Web Conference 2026 (pp. 8008-8016).
29. 29. Ou, L., Liu, Y., Li, Z., Bai, X., & Peng, Y. (2026, March). Structure-Grounded Training Strategies Aid Generalization in Stereo Matching. In 2026 International Conference on 3D Vision (3DV) (pp. 125-134). IEEE.
30. 30. Tang, J., Yuan, X., Liu, J., Yu, J., Dong, X., Chen, L., ... & Yin, J. (2026, August). Sake: Self-aware knowledge exploitation-exploration for grounded multimodal named entity recognition. In Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2 (pp. 4592-4603).
31. 31. Girshick, R. (2015). Fast R-CNN. In Proceedings of ICCV.
32. 32. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., & Joulin, A. (2021). Emerging properties in self-supervised vision transformers. In Proceedings of ICCV.
33. 33. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2016). Rethinking the Inception architecture for computer vision. In Proceedings of CVPR.
34. 34. Tao, J., Lyu, R., & Cao, X. (2026). A Deep Learning-Based Automated Content Moderation Framework for Online Platforms.Future-Adaptive Intelligence and Lifelong Systems,1(1).
35. 35. Wang, Y., & Ling, C. (2025). Comparing SAS® and R Approaches in Reshaping data. In PharmaSUG 2025 Conference Proceedings.
36. 36. Zhang, Y., Chen, X., Zhao, L., Ji, Y., Liu, P., & Cheung, Y. M. (2025). Online heterogeneous feature selection. IEEE Transactions on Cybernetics.