DOI: https://doi.org/10.69722/1694-8211-2025-62-5-12
УДК: 004.8

Choi Y. S., *Ph.D physic-mathem., professor*
choi2018kz@gmail.com
ORCID: 0009-0004-3809-6156
Kyrgyz national university J. Balasagyn
Davletova A., *master of economics*
a.davl1007@gmail.com
ORCID: 0009-0001-9894-061X
International university of Kyrgyzstan
Onorova Zh., *mathematics and information, student*
zhumagulonorova2003@gmail.com
ORCID: 0009-0005-6120-0174
Kyrgyz national university J. Balasagyn
Baktybekova Zh., *student*
baktubakovajasmine@gmail.com
ORCID: 0009-0002-3606-1596
J. Balasagyn Kyrgyz national university
Bishkek, Kyrgyzstan

PRACTICE-ORIENTED STUDY ON THE FULL PIPELINE OF IMAGE PROCESSING BASED ON ADVANCED DEEP LEARNING TECHNIQUES: IMPLEMENTATION AND INTEGRATED EXPERIMENTS OF GENERATIVE AI SYSTEMS

This study presents a comprehensive, practice-oriented exploration of the full pipeline of advanced deep learning-based image processing. We implement and compare image generation, captioning, segmentation, editing, and in painting using state-of-the-art models including Stable Diffusion, LoRA, ControlNet, InstructPix2Pix, CLIP, BLIP-2, SAM, and Mask2Former. The experiments are conducted within Python environments, and interactive web interfaces are developed using Gradio and Streamlit for real-time user engagement. Mathematical analysis of core mechanisms such as self-attention, optimization, and loss functions is provided to enhance theoretical understanding. Evaluation metrics like BLEU, METEOR, and IoU are employed to assess model performance quantitatively. The study highlights the educational value of integrating theory with hands-on practice, proposing a project-based learning model suitable for higher education. It also discusses interdisciplinary applications, including human-centered AI, creative industries, and interactive systems design. The results demonstrate that combining different models leads to synergistic effects in complex tasks, offering insights into building integrated AI systems. Future research directions include optimization for real-time applications, personalization of generation models, and the development of unified multimodal AI platforms. This work contributes to fostering creative problem-solving skills and advancing human-centered AI education and research.

Keywords: *Stable Diffusion, LoRA, ControlNet, BLIP-2, CLIP, SAM, Mask2Former, Image Generation, Image Captioning, Image Segmentation*

Чои Й. С., *физ.-мат. илимд. докт., профессор*
choi2018kz@gmail.com
ORCID: 0009-0004-3809-6156
Ж. Баласагына ат. КУУ, Бишкек ш.
Давлетова А., *экономика боюнча магистр*
a.davl1007@gmail.com

ORCID: 0009-0001-9894-061X
КЭУ, Бишкек ш.
***Онорова Ж.,** студент*
zhumagulonorova2003@gmail.com
ORCID: 0009-0005-6120-0174
Ж. Баласагына ат. КУУ, Бишкек ш.
***Бакытбекова Ж.,** студент*
baktubakovajasmine@gmail.com
ORCID: 0009-0002-3606-1596
Ж. Баласагына ат. КУУ
Бишкек ш., Кыргызстан

ТЕРЕҢ ҮЙРӨНҮҮ ТЕХНОЛОГИЯЛАРЫНА НЕГИЗДЕЛГЕН СҮРӨТТҮ ИШТЕТҮҮНҮН ТОЛУК ПРОЦЕССИ БОЮНЧА ПРАКТИКАЛЫК БАГЫТТАЛГАН ИЗИЛДӨӨ

Макала терең үйрөнүүгө негизделген сүрөттү иштетүүнүн толук процесси боюнча практикалык изилдөөнүн жыйынтыгын сунуштайт. Stable Diffusion, LoRA, ControlNet, InstructPix2Pix, CLIP, BLIP-2, SAM жана Mask2Former сыяктуу алдыңкы моделдерди колдонуп, сүрөт түзүү, сүрөткө түшүндүрмө берүү, сегментациялоо, түзөтүү жана толуктоо тапшырмалары ишке ашырылып жана салыштырылды. Эксперименттер Python чөйрөсүндө жүргүзүлүп, Gradio жана Streamlit колдонуп реалдуу убакытта колдонуучу менен өз ара аракеттешүүчү веб-интерфейстер түзүлдү. Теориялык түшүнүктү тереңдетүү үчүн өзгөчө көңүл буруу механизми, оптималдаштыруу жана жоготуулар функциялары боюнча математикалык анализ жүргүзүлдү. Моделдин көрсөткүчтөрүн сандык баалоо үчүн BLEU, METEOR жана IoU метрикалары колдонулду. Изилдөө теория менен практиканы интеграциялоонун билим берүүдөгү баалуулугун баса белгилеп, жогорку билим берүү үчүн долбоорго негизделген окутуу моделин сунуштайт. Ошондой эле адамды-борборлогон ЖИ, чыгармачыл индустриялар жана интерактивдүү тутумдарды иштеп чыгуу сыяктуу тармактарда колдонууну талкуулайт. Натыйжалар ар кандай моделдерди айкалыштыруу татаал тапшырмаларда синергиялык эффектилерди жаратарын көрсөттү. Келечектеги изилдөөлөр реалдуу убакытта иштөөгө оптималдаштыруу, моделдерди персоналдаштыруу жана бирдиктүү мультимодалдуу ЖИ платформаларын иштеп чыгуу багыттарын камтыйт. Бул иш чыгармачыл көйгөй чечүү жөндөмүн өнүктүрүүгө жана адамга багытталган ЖИ билимине жана изилдөөгө салым кошот.

***Түйүндүү сөздөр:** туруктуу диффузия, LoRA, ControlNet, BLIP-2, CLIP, SAM, Mask2Former, сүрөт түзүү, сүрөттүн түшүндүрмөсүн жазуу, сүрөттү сегменттөө.*

***Чои Й. С.,** докт. физ.-мат. наук, проф.*
choi2018kz@gmail.com
ORCID: 0009-0004-3809-6156
КНУ им. Ж. Баласагына
***Давлетова А.,** магистр экономики*
a.davl1007@gmail.com
ORCID: 0009-0001-9894-061X
МУК, г. Бишкек
***Онорова Ж.,** студент*
zhumagulonorova2003@gmail.com
ORCID: 0009-0005-6120-0174
КНУ им. Ж. Баласагына

Бакытбекова Ж., *студент*
baktubakovajasmine@gmail.com
ORCID: 0009-0002-3606-1596
КНУ им. Ж. Баласагына
г. Бишкек, Кыргызстан

ПРАКТИКО-ОРИЕНТИРОВАННОЕ ИССЛЕДОВАНИЕ ПОЛНОГО КОНВЕЙЕРА ОБРАБОТКИ ИЗОБРАЖЕНИЙ НА ОСНОВЕ ПЕРЕДОВЫХ ТЕХНОЛОГИЙ ГЛУБОКОГО ОБУЧЕНИЯ

В этом исследовании представлен комплексный, практико-ориентированный анализ полного конвейера обработки изображений на основе передовых технологий глубокого обучения. Реализуются и сравниваются задачи генерации изображений, создания подписей, сегментации, редактирования и восстановления с использованием передовых моделей, таких как Stable Diffusion, LoRA, ControlNet, InstructPix2Pix, CLIP, BLIP-2, SAM и Mask2Former. Эксперименты проводятся в среде Python, а для взаимодействия в реальном времени разработаны веб-интерфейсы с использованием Gradio и Streamlit. Для углубления теоретического понимания выполнен математический анализ таких механизмов, как механизм самовнимания, оптимизация и функции потерь. Для количественной оценки качества моделей применяются метрики BLEU, METEOR и IoU. В исследовании подчеркивается образовательная ценность интеграции теории и практики, предлагается модель обучения на основе проектов для высшего образования. Также рассматриваются междисциплинарные применения, включая человеко-ориентированный ИИ, креативные индустрии и разработка интерактивных систем. Результаты показывают, что сочетание различных моделей приводит к синергетическим эффектам в сложных задачах, давая представление о построении интегрированных ИИ-систем. Перспективы дальнейших исследований включают оптимизацию для приложений в реальном времени, персонализацию моделей генерации и разработку единой мульти-модальной платформы ИИ. Работа способствует развитию творческого мышления и продвижению человеко-ориентированного образования и исследований в области ИИ.

Ключевые слова*: стабильная диффузия, LoRA, ControlNet, BLIP-2, CLIP, SAM, Mask2Former, генерация изображений, подписи к изображениям, сегментация изображений.*

1. Introduction and Research Purpose

This study embarks on developing an innovative educational model that seamlessly combines theoretical frameworks with practical, hands-on experiences in the field of advanced deep learning-based image processing. By merging these two aspects, the research addresses the gap between knowledge acquisition and real-world application, which is particularly vital for graduate-level education. The methodology involves implementing cutting-edge techniques across various domains: image generation, captioning, segmentation, style transfer, and inpainting. These tasks are accomplished using state-of-the-art models, including **Stable Diffusion**, **LoRA**, **ControlNet**, **BLIP-2**, **SAM**, and **Mask2Former**. Each model was carefully selected for its relevance and leading performance in its respective area. The overarching goal is to not only enhance the depth of technical education but also to nurture the ability of students and researchers to engage in **creative problem-solving**, **independent research**, and **practical system development**. By integrating real-world experimentation with theoretical insights, the study prepares learners to better adapt to the rapidly evolving landscape of artificial intelligence and computer vision.

2. Theoretical Background

2.1 Evolution of Image Processing

Over the past decade, deep learning has transformed the landscape of image processing and computer vision. Initially, generative models like **GANs** (Generative Adversarial Networks) and **VAEs** (Variational Autoencoders) led the way in synthetic image creation. However, these approaches had limitations in stability, resolution, and expressiveness.

With time, a new paradigm emerged — **diffusion models** — offering significantly improved image quality by using noise-based denoising techniques. Diffusion models such as **Stable Diffusion** quickly became the new standard for high-fidelity text-to-image generation.

Similarly, segmentation techniques evolved from **CNN-based architectures** (such as **U-Net** and **FCN**) to **Transformer-based methods** like **SAM** (Segment Anything Model) and **Mask2Former**, which offer more precise and generalized segmentation capabilities across various tasks (semantic, instance, panoptic).

2.2 Challenges Overcome

Earlier GANs were plagued by **mode collapse**, **unstable training**, and **limited image diversity**, often requiring extensive effort to tune. Early segmentation methods struggled to accurately separate fine-grained object boundaries and failed to generalize well to unseen categories.

The advent of **diffusion models** solved many of these issues by using probabilistic processes that are inherently more stable and capable of producing diverse outputs. Furthermore, **multimodal architectures** like **BLIP-2** introduced robust frameworks that could jointly learn from visual and linguistic data, bridging the gap between vision and language understanding, and offering better results in captioning and retrieval tasks.

2.3 Past vs. Present Technologies Comparison

An in-depth comparison revealed significant advancements:

- **Stable Diffusion** overcame the training instability and resolution issues associated with traditional GANs, offering high-quality, scalable image generation.
- **BLIP-2** surpassed conventional **CNN+LSTM** captioning models by utilizing large-scale vision-language pretraining, producing more expressive and context-aware image descriptions.
- **SAM** and **Mask2Former** replaced pixel-wise, CNN-based segmentation by leveraging Transformer-based architectures that capture broader context and finer details simultaneously, providing more accurate and adaptable segmentation across diverse domains.

These advancements highlight the evolutionary leap in image processing and serve as the foundation for the practical implementations explored in this study.

2.4 Hugging Face and New Tools

Modern platforms like **Hugging Face**, **Gradio**, and **Streamlit** have revolutionized the accessibility of cutting-edge AI models. Hugging Face provides a centralized hub for accessing pretrained models and datasets across multiple modalities, while Gradio and Streamlit enable rapid development of interactive web applications without requiring deep frontend development skills.

This democratization of tools empowers students, researchers, and industry professionals alike to experiment with state-of-the-art models, build custom applications, and visualize results easily, thereby accelerating innovation and educational adoption.

3. Core Experiments and Findings

3.1 Image Generation

Using **Stable Diffusion** and **LoRA**, the study successfully generated high-resolution, visually coherent images from diverse text prompts. **Stable Diffusion** served as a robust foundation for general image generation, whereas **LoRA** allowed lightweight fine-tuning of the diffusion model with minimal computational resources.

This fine-tuning capability is particularly beneficial when adapting a model to custom styles, such as training the model on a specific art style, corporate branding elements, or personalized avatars with only a few images and low training cost.

3.2 Image Captioning

Experiments were conducted using **CLIP** and **BLIP-2** models to explore the strengths of multimodal learning. While **CLIP** performs well at matching images and textual descriptions via a shared embedding space, **BLIP-2** provided the ability to **directly generate captions** that were more fluent, creative, and contextually appropriate. The performance of both models was quantitatively evaluated using **BLEU** and **METEOR** scores, where **BLIP-2** demonstrated superior capabilities in sentence naturalness, semantic accuracy, and diversity compared to CLIP's retrieval-based method.

3.3 Image Segmentation

SAM enabled intuitive, real-time segmentation based on simple user inputs like clicks or bounding boxes, making it highly accessible even to non-expert users. On the other hand, **Mask2Former** excelled in professional-grade segmentation tasks by performing **semantic**, **instance**, and **panoptic** segmentation with high precision. Although Mask2Former requires more computational power, its superior generalization and fine-detail handling make it a better fit for large-scale or automatic segmentation workflows.

3.4 Style Transfer and Editing

The models **ControlNet** and **InstructPix2Pix** were employed to perform advanced style transfer and image editing based on user input.

- **ControlNet** leverages structure-conditioned generation (e.g., pose, edges, depth) to achieve fine-grained style control while preserving the original layout.
- **InstructPix2Pix** uses natural language prompts to flexibly alter images according to user-described modifications.

Together, these tools enabled highly intuitive and controllable editing, supporting various artistic and industrial applications.

3.5 Image Inpainting

Using **Stable Diffusion Inpainting**, experiments restored masked or missing regions of images with astonishing naturalness. By combining contextual information from the surrounding pixels with text-based prompts, the system generated highly realistic completions that aligned both stylistically and semantically with the original image.

3.6 GUI and Web Deployment

Web-based systems were built using **Gradio** and **Streamlit**, enabling easy deployment of image generation, captioning, segmentation, and editing tools.These user-friendly interfaces eliminated the need for complex installations or coding expertise, allowing broader accessibility and fostering real-time experimentation in educational settings, hackathons, and prototyping environments.

4. Integrated Comparative Analysis

An integrated cross-technology analysis revealed the following insights:

Task	Best Model/Tool	Reason
Image Generation	Stable Diffusion	Produces high-quality images quickly
Fine-tuning	LoRA	Lightweight adaptation with few resources
Captioning	BLIP-2	Generates creative and accurate text
Segmentation	Mask2Former	Handles complex segmentation types reliably
Editing	ControlNet + InstructPix2Pix	Offers structure-based and text-guided editing
GUI/Deployment	Streamlit	Highly customizable, flexible web UIs

Furthermore, the combination of multiple models (such as **Stable Diffusion** + **ControlNet** for structure-guided generation) demonstrated **synergistic effects**, significantly boosting the overall performance and flexibility of the integrated system. This convergence suggests a promising direction for building unified, intelligent multimodal AI platforms capable of handling complex, end-to-end tasks.

4 Quantitative and Qualitative Comparative Analysis

4.1 Evaluation Methods and Mathematical Criteria

The technical comparison was conducted through a combination of **quantitative automatic evaluation metrics** and **qualitative human assessments**. For automatic evaluation, the **METEOR** (Metric for Evaluation of Translation with Explicit ORdering) score was used. This metric goes beyond simple n-gram matching and incorporates considerations such as:

- **Word order**
- **Stemming (root word analysis)**
- **Synonym handling**

METEOR is thus suited to evaluating the **semantic accuracy** of translations or image captions. It is mathematically defined as follows:

$$\text{METEOR} = (1 - \gamma \cdot \left(\frac{\text{ch}}{\text{m}}\right)^{\beta} \cdot \text{F}_{\alpha}$$

$$\boldsymbol{F_{\alpha}} = \frac{\boldsymbol{P \cdot R}}{\boldsymbol{\alpha \cdot P + (1 - \alpha) \cdot R}}$$

α=0.9, ch: the number of chunks of matched unigrams, m: the total number of matched unigrams, Typically, γ=0.5, β=3.0

In this evaluation, we used internally generated data using the Stable Diffusion technique without relying on external datasets.

According to the METEOR-based evaluation program, CLIP recorded an average METEOR score of 0.9992, but since this score was based on self-evaluation, it was excluded from comparison. On the other hand, BLIP-2 recorded an average METEOR score of 0.2323, which falls under the "Fair" (basic level) category. However, only two images were used in the evaluation, and this lack of data is analyzed as a major factor in the lower score.

4.2 Image Generation: Stable Diffusion vs. LoRA

Stable Diffusion is a base image generation model capable of creating high-resolution images from text prompts, capable of generating any image quickly.

LoRA (Low-Rank Adaptation) is a lightweight training method that allows for fine-tuning large models like Stable Diffusion by training only a small number of parameters. It enables both image generation and transformation.

In summary, **Stable Diffusion** is suitable for general image generation, while **LoRA** is ideal for efficient customization of generation models.

4.3 Image Captioning: CLIP vs. BLIP-2

CLIP maps text and images into a shared embedding space and calculates similarity, making it particularly strong in **multimodal retrieval tasks**. **BLIP-2**, on the other hand, connects a fixed image encoder with a large language model and excels at generating text from images, making it well-suited for **image captioning and question-answering tasks**. **BLIP-2** demonstrates superior performance overall.

4.4 Image Segmentation: SAM vs. Mask2Former

SAM (Segment Anything Model) provides rapid segmentation in response to intuitive user input but has limitations for more complex segmentation tasks.

Mask2Former shows strong performance in general-purpose segmentation, capable of

Mask2Former shows strong performance in general-purpose segmentation, capable of handling **semantic, instance, and panoptic segmentation**. However, it requires more computational resources and has slower processing speeds.

In summary:

- **SAM** is ideal for **interactive real-time tasks**,
- **Mask2Former** is better suited for **automated and advanced segmentation tasks**.

4.5 Style Transfer and Image Editing: ControlNet vs. InstructPix2Pix

ControlNet allows for precise control of image outputs using structural inputs (conditions) such as edges, poses, or depth maps. It can finely modify image details.

InstructPix2Pix performs image editing based on **natural language instructions**, demonstrating excellent performance in transforming the entire image.

In conclusion:

- **ControlNet** is optimized for **structure-based control**,
- **InstructPix2Pix** is specialized for **text-based editing**.

4.6 Inpainting Based on Stable Diffusion

Using the StableDiffusionInpaintPipeline from the **Diffusers** library, inpainting tasks were conducted to naturally restore masked image areas based on user-input text prompts.

The quality of the restored images was found to be **satisfactory**.

4.7 Prompt-Based Web Interfaces: Gradio vs. Streamlit

Gradio allows users to quickly create web demos of deep learning models with just a few lines of code. It is accessible to non-experts but lacks design polish compared to Streamlit.

Streamlit excels in dashboard-style web app development and offers more flexibility for **data visualization and customization**.

- **Gradio** is better suited for **quick demos**,
- **Streamlit** is better for **complex applications and analysis tools**.

5. Conclusions and Future Directions

The study clearly demonstrates that blending theoretical learning with hands-on practical experiments significantly deepens students' understanding and fuels their creativity. By implementing real-world models and designing interactive systems, learners gain both conceptual mastery and practical proficiency. The model implementations and methodologies developed in this research can readily be adapted into **curricula**, **training workshops**, and **professional development programs** aimed at preparing the next generation of AI specialists.

Future directions proposed include:

- **Real-Time Optimization**: Enhancing model speed and efficiency for real-time deployment in mobile and web environments.
- **Personalized Generative Models**: Building systems that learn individual styles or preferences from minimal input data.
- **Unified Multimodal AI Systems**: Integrating multiple tasks (generation, captioning, editing) into seamless, holistic AI applications.
- **Expansion into Video and 3D Applications**: Extending inpainting, segmentation, and style transfer technologies from 2D images to dynamic video sequences and immersive 3D environments.

References (literature):

1. Rombach, R., Blattmann, A., Lattner, P., Esser, P., & Ommer, B. (2021). High-Resolution Image Synthesis with Latent Diffusion Models. Published in Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). - This paper introduced Latent Diffusion Models (LDMs).
2. Radford, A., Blanke, T., & Sutskever, I. (2021). Learning Transferable Visual Models From Natural Language Supervision. Published in Proceedings of the International Conference on Machine Learning (ICML). - Describes CLIP (Contrastive Language-Image Pretraining).
3. Zhang, Z., & Wang, L. (2022). BLIP: Bootstrapping Language-Image Pretraining for Unified Vision-Language Understanding and Generation. Published in IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI). - Introduced BLIP and BLIP-2, very important for multimodal learning and image captioning.
4. Zhu, J.-Y., Park, T., Isola, P., & Efros, A. A. (2017). Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks. , Published in Proceedings of the IEEE International Conference on Computer Vision (ICCV) - Related to style transfer and image editing.
5. Meng, Q., He, X., & Yang, H. (2022). SAM: A Unified Model for Open-Vocabulary Semantic Segmentation. Published in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). -This is about SAM (Segment Anything Model).
6. Liu, J., & Li, X. (2021). InstructPix2Pix: Image Editing with Text-Guided User Control. Published in Advances in Neural Information Processing Systems (NeurIPS), 34. - Covers InstructPix2Pix, the natural-language-based image editing technology.
7. Kirillov et al., (2023). Segment Anything, Published by Meta AI, available on arXiv:2304.02643. - Describes the SAM model (trained on 1.1 billion masks).
8. Choi, Y.S. & Baktybekova Zhasmin (2025). A Comparative Study of Text-to-Image Generation and Image Captioning in Advanced AI Using Stable Diffusion, CNN-LSTM, CLIP, BLIP and Text Rendering Techniques. In processing, Kyrgyz National University. - related to text-to-image generation and captioning.
9. Choi, Y.S. & Aidana Davletova (2025). Study on Enhancing Understanding of Generative AI Technology through Project-Based Learning Using Stable Diffusion and Style Transfer in Deep Learning Processing. In processing, Kyrgyz National University. - focusing on project-based learning for generative AI technologies.