# *CoTextor*: Training-Free Modular Multilingual Text Editing via Layered Disentanglement and Depth-Aware Fusion

**Zhenyu Yu**[2], **Mohd. Yamani Idna Idris**[2], **Pei Wang**[1,*], **Rizwan Qureshi**[3]
[1]Kunming University of Science and Technology
[2]Universiti Malaya
[3]University of Central Florida
yuzhenyuyxl@foxmail.com; yamani@um.edu.my; peiwang@kust.edu.cn;
rrrizwan2-c@my.cityu.edu.hk

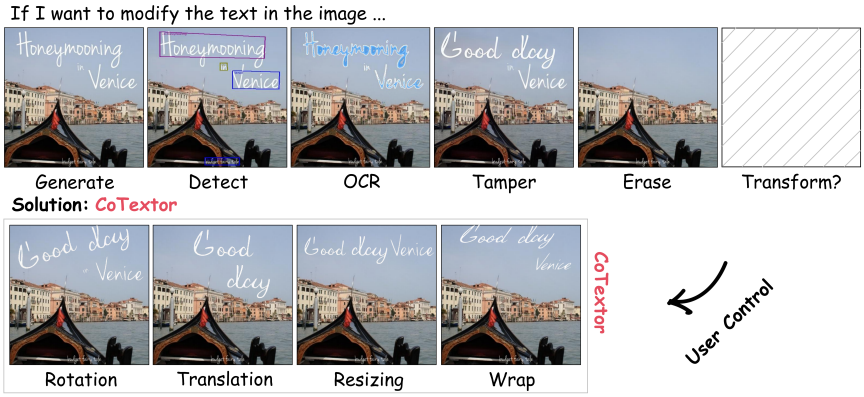


Figure 1: Motivation. While prior work addresses text generation, detection, tampering, and removal, *controllable multilingual text editing*—with layout preservation and visual consistency—remains largely unsolved. *CoTextor* fills this gap by offering a modular, depth-aware framework for human-guided, reversible visual transformation.

## Abstract

We introduce *CoTextor*, a modular and training-free framework for multilingual text editing in images, designed to support human-AI co-creation through a user-controllable and reversible workflow. Unlike diffusion-based systems that operate as black boxes, *CoTextor* separates the editing process into transparent layers—foreground extraction, background inpainting, semantic rewriting, and depth-aware reintegration—allowing precise user-guided operations such as rotation, translation, scaling, and warping. To ensure realism, we introduce a perceptually guided integration module that enhances photometric and geometric coherence during text reinsertion. Built entirely from publicly available pretrained components, *CoTextor* is accessible to non-technical, multilingual users, requiring no retraining or annotation. Through real-world scenarios in poster localization, street art remixing, and educational content creation, we demonstrate how *CoTextor* enables inclusive and expressive visual storytelling across cultural and linguistic contexts.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Creative AI Track.

# 1 Introduction

Text in images plays a multifaceted role—it communicates information, anchors visual narratives, facilitates cross-cultural understanding, and often embodies aesthetic and cultural intent [31, 3, 57, 47, 25, 52, 67]. In real-world creative practices—such as poster localization, comic adaptation, or street art remixing—text is not merely rendered, but reimagined through design-driven manipulation [22, 8, 60, 44, 24, 51]. Artists and educators frequently reshape semantics, reposition layout, or stylize embedded language to suit audience, format, and intent. These tasks often require not only linguistic fluency but also visual sensitivity, especially when communicating across cultures and media formats [56, 59, 15, 65, 30]. Moreover, such editing is rarely a one-step operation—it typically involves iterative revision, feedback, and adjustment, especially in collaborative or educational settings [21, 7, 58, 54, 64, 29].

Recent progress in generative AI, especially diffusion-based models [40, 43, 48, 38, 7, 66], has improved multilingual rendering and visual synthesis. However, these models often function as opaque generators with limited spatial control and no support for post-editing. Their one-shot, prompt-only design makes them ill-suited for iterative workflows, layout preservation, or collaborative human-AI co-creation [4, 63, 33, 12]. While impressive in image realism, these models lack the transparency and interactivity needed in real-world authoring scenarios. As a result, creators are often forced to trade expressiveness and precision for automation—especially in multilingual or design-sensitive tasks (see Figure 1).

In this work, we propose ***CoTextor***, a modular and *training-free* framework that enables controllable multilingual text editing in images. Instead of optimizing for generation fidelity or new model architectures [53], *CoTextor* focuses on usability, transparency, and creative control. It decomposes the editing process into interpretable modules: text extraction, background reconstruction, prompt-guided rewriting, and depth-aware reintegration. Each module exposes clear control points to the user, allowing them to intervene at any stage of the workflow without retraining or coding effort. This design supports not only technical modularity, but also creative flexibility, allowing human decisions to shape the outcome dynamically and deliberately [21, 26]. *CoTextor* is designed to empower users—especially non-technical, multilingual, or culturally diverse creators. By supporting layout-aware control and script-sensitive rendering, it allows human intent to guide the system at every stage. Unlike conventional systems that center model output, *CoTextor* centers human authorship—prioritizing reversibility, customization, and co-creative expression [16, 11, 28]. This shift from prompt-driven generation to editable, interpretable workflows broadens the accessibility of creative AI tools. Through case studies in multilingual design, visual remixing, and educational content creation, we show how *CoTextor* contributes to the broader vision of Creative AI: not just automating outputs, but augmenting authorship [50, 5, 27].

**Our main contributions are:** (1) **A user-controllable editing workflow.** We present *CoTextor*, a modular and training-free system that supports human-in-the-loop multilingual text editing through a layered, reversible pipeline. (2) **Perceptually guided integration.** We employ a lightweight, depth-aware composition module that enhances photometric and geometric coherence, supporting realism in human-authored visual narratives. (3) **Culturally inclusive and accessible design.** *CoTextor* enables creative users—including non-technical and multilingual audiences—to perform expressive transformations without code, annotations, or retraining.

# 2 Related Work

## 2.1 Generative Models for Text-in-Image Editing

Diffusion-based models have become foundational in generative vision pipelines, enabling high-quality text-to-image synthesis from prompts [40, 41, 43, 37, 36]. Models like DALL·E and Stable Diffusion [49] inspired follow-up work targeting embedded text rendering, including AnyText [48] and TextDiffuser [8], which integrate OCR [23, 13, 34, 7]signals and stylized rendering for multilingual scenarios. However, these systems are typically optimized for one-shot generation rather than iterative or user-guided editing. Their architectures prioritize visual plausibility over spatial controllability or layout preservation, making them less suitable for workflows involving back-and-forth revision, geometric reasoning, or narrative adaptation [4, 35, 9]. For instance, once a diffusion-based result is generated, there is often no native way to adjust a specific text element without rerunning

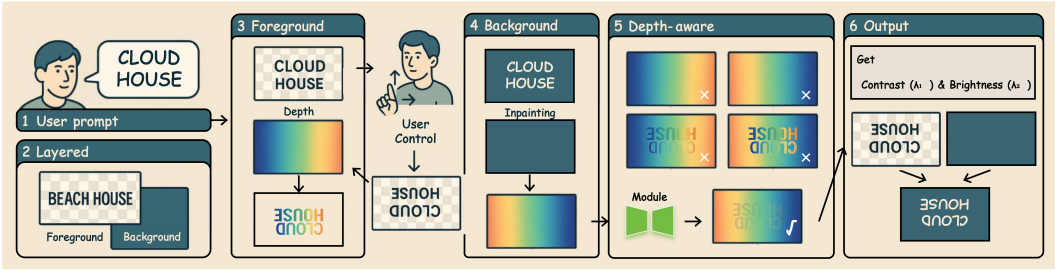


Figure 2: Overview of the *CoTextor* framework. The pipeline consists: (1) OCR- and SAM-based foreground extraction; (2) Background inpainting after text removal; (3) User-control geometric transformation of the text layer; (4) Depth-aware composition to ensure realistic integration.

the full model. Some recent works such as Muse [6] and RePaint [32] attempt inpainting-based regeneration, but still offer limited textual consistency and geometric control. Crucially, these tools position users as prompt writers, not as co-creators [63].

### 2.2 Controllability and Human-Guided Editing

Recent efforts have explored controllable editing interfaces, particularly to expand the creative scope of generative models. For instance, GlyphControl [53] enables glyph-aware style transfer, and InstructPix2Pix [4] supports high-level region editing via natural language. TextDiffuser [8] incorporates segmentation masks to constrain placement, but remains prompt-bound and non-reversible.

Other systems such as DreamBooth [42] and ControlNet [63] introduce conditioning strategies to guide generation via keypoints, depth maps, or sketches. While these enhance controllability to some extent, they still operate on fixed generative flows and do not expose editable layers or iterative workflows. In most cases, the user has limited agency beyond initial conditioning. These approaches still treat editing as a downstream task to generation, rather than designing editing-first systems. They often lack support for visual feedback, explicit undo, or bidirectional control—key elements in any human-centered creative process [21, 10, 61]. Our work builds on these ideas, but repositions control as the starting point, not the afterthought.

### 2.3 Modular Design and Disentangled Creativity

Foreground-background disentanglement has long been useful in image editing [1, 55]. Modern inpainting models such as LaMa [45] improve photorealistic filling, but typically treat foreground and background as stylistic artifacts rather than semantically distinct elements.

In the context of text editing, the semantic-visual duality of glyphs—both symbolic and graphical—requires more structured separation. A few recent works such as DeepRemaster [18] and SEAN [68] begin to explore layer-wise disentanglement for style manipulation, though not in the context of interactive textual editing. *CoTextor* adopts this principle not only as an architectural convenience but as a creative affordance. By modularizing the process—text recognition, region selection, tampering, and recomposition—it supports reversible, interpretable, and multilingual editing. More importantly, this design foregrounds human agency, allowing creators to guide not just what is edited, but how and why [50, 62].

## 3 Human-AI Co-Creation Workflow

*CoTextor* is designed not as a monolithic generation engine, but as a modular editing framework that supports human-AI co-creation. Its architecture encourages active user participation at every stage, enabling a collaborative interplay between human intent and algorithmic support in shaping visual outcomes (see Figure 2 and Figure 3(a)).

**Stage 1: Interactive Foreground Extraction.** Users begin by identifying one or more text regions in the image. An OCR model (e.g., EasyOCR [19]) detects all textual elements, after which the user selects which content to edit. *CoTextor* applies the Segment Anything Model (SAM) [20] for initial

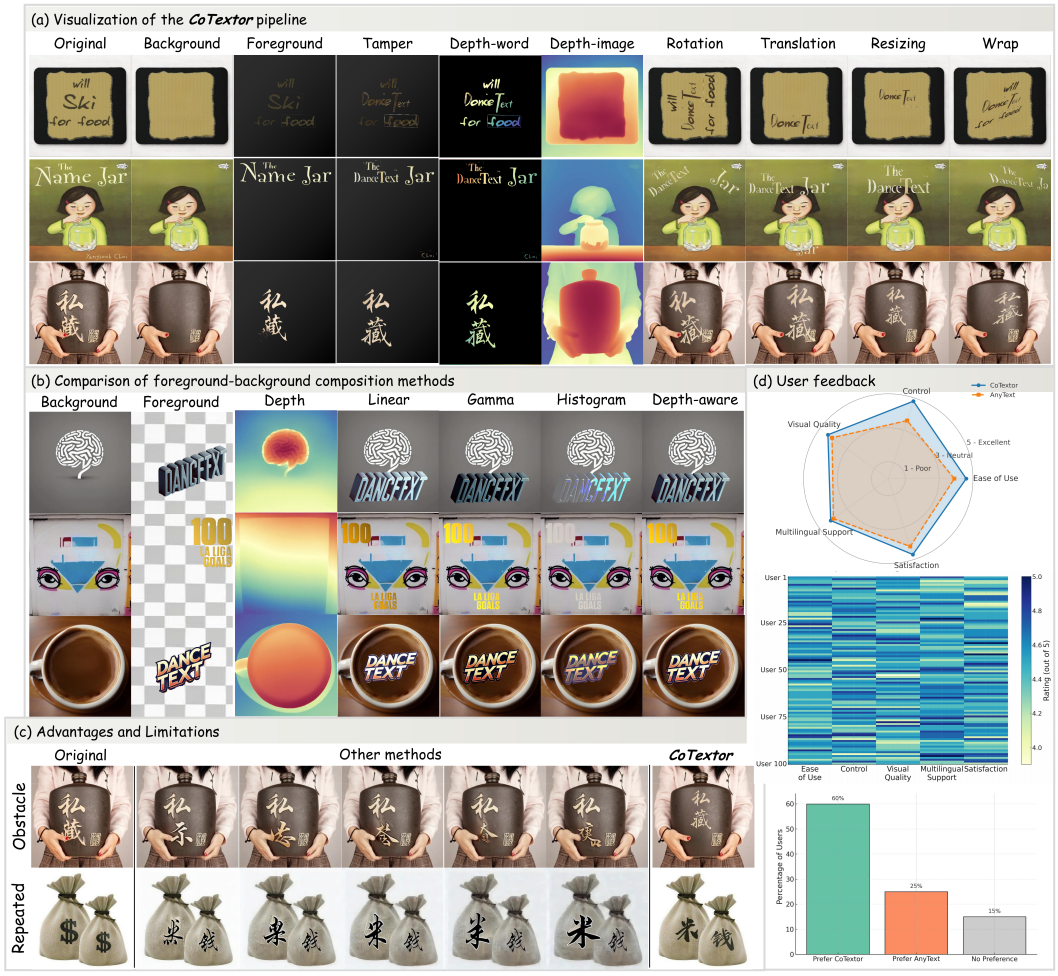


Figure 3: *CoTextor* framework and evaluation overview. (a) Visual pipeline. (b) Composition quality comparison. (c) Performance under challenging scenarios. (d) User study summary.

segmentation, followed by $k$-Means clustering to remove residual noise. This results in a clean and editable foreground layer, giving users direct control over *what* to modify.

**Stage 2: Clean Background Reconstruction.** After text removal, the background is automatically reconstructed to preserve visual context. A pretrained inpainting model (e.g., LaMa [45]) synthesizes plausible textures in masked regions, producing a scene-consistent canvas on which new edits can be placed. This enables fluid repositioning without breaking layout harmony.

**Stage 3: Prompt-Guided Tampering and Geometric Transformation.** Users rewrite selected text through multilingual prompts using AnyText, and apply geometric transformations—rotation, translation, scaling, warping—via intuitive controls. Unlike prompt-only models, *CoTextor* supports real-time visual feedback, empowering users to iteratively refine layout, semantics, and style. This stage prioritizes creative exploration and non-linear editing.

**Stage 4: Depth-Aware Reintegration.** To maintain perceptual realism, *CoTextor* estimates depth maps for both the inpainted background and the transformed foreground. A depth-aware fusion module then adjusts contrast and brightness to align with local lighting and geometry. This ensures the edited content integrates seamlessly, even under complex scene conditions.

By disentangling recognition, manipulation, and composition, *CoTextor* enables expressive visual editing that is interactive, reversible, and culturally adaptable. Its modular structure foregrounds user agency—allowing creators not just to generate content, but to actively author visual meaning across languages and contexts.

## 4 Implementation Details

*CoTextor* is built entirely from publicly available pretrained models. We evaluate *CoTextor* on real-world samples drawn from the AnyWord-3M benchmark [48]. Metrics include Structural Accuracy (SA) [48], Normalized Edit Distance (NED) [48], Fréchet Inception Distance (FID) [17], Bhattacharyya Coefficient (BC) [2], Chi-Square Distance (CS) [14], Correlation (Corr) [39], and Histogram Intersection (Inter) [46]. All experiments were conducted on a single NVIDIA RTX 4090 GPU. The full implementation—including mathematical formulations and ablation settings is provided in Appendix A to support transparency and reproducibility.

## 5 Evaluation: Human-Centered Visual Editing

### 5.1 Qualitative Use Cases: Human-AI Co-Creation in Practice

To assess the practical impact of *CoTextor* in real-world creative workflows, we present three illustrative use cases involving multilingual design, public expression, and educational adaptation. These scenarios demonstrate how modularity, reversibility, and spatial control enable users to co-author expressive visual content across cultural and linguistic contexts.

**Multilingual Poster Redesign.** A visual designer begins with an English-language promotional poster and aims to localize it into Arabic and Chinese. *CoTextor* enables selective text extraction, prompt-based rewriting via AnyText, and flexible layout adjustment through geometric transformation. The designer is not constrained by prompt syntax or static outputs but actively reshapes the composition to fit new audiences. The depth-aware fusion module ensures that the localized text blends with the scene's lighting and perspective—an essential need in culturally adaptive visual storytelling.

**Street Art Remix.** In a reinterpretation of urban graffiti, a user selects stylized text on a brick wall and transforms the message (e.g., from `I love you` to `I hate you`), while retaining its style and curvature. *CoTextor* supports real-time warping and lighting-aware rendering, allowing users to reimagine public messages without disrupting scene integrity. Unlike black-box pipelines, *CoTextor* gives users control over both message and form—a key factor in creative reappropriation.

**Educational Storybook Translation.** An educator localizing visual materials for multilingual learners replaces in-image labels such as `start` or `help`. *CoTextor* preserves the surrounding illustration while enabling precise word-level substitution. With reversible editing and previewable iterations, educators can ensure clarity and consistency across language versions, making it easier to support diverse learning communities.

These cases highlight *CoTextor*'s ability to support visual authorship that is not only technically plausible but also socially and culturally responsive. The user is repositioned not as a prompt-giver but as a co-creator, shaping narrative, layout, and linguistic voice.

### 5.2 Perceptual Consistency Analysis

We evaluate CoTextor's depth-aware fusion against standard composition methods—linear blending, gamma correction, and histogram matching—using four perceptual metrics: brightness consistency (BC), color shift (CS), semantic correlation (Corr), and interaction coherence (Inter) (see Figure 3(b), Table 1). CoTextor achieves the highest Corr (0.9915) and Inter (1.8251), indicating better semantic alignment and scene integration. Although histogram matching yields better BC, it underperforms in semantic fidelity. These results confirm the effectiveness of depth-guided adjustment for perceptually coherent and trustworthy text editing. Additional quantitative comparisons are provided in Appendix D.

### 5.3 User Study and Preference Analysis

To complement our quantitative metrics, we conducted a small-scale user study involving 30 participants with backgrounds in visual design, language education, and digital media. Each participant was asked to complete editing tasks using both CoTextor and a baseline diffusion model. As shown in Figure 3(d), users rated CoTextor higher in controllability (mean: 4.6/5), visual quality (4.5), and overall satisfaction (4.7), citing its modular workflow and real-time feedback as key strengths.

## 6 Discussion

Table 1: Ablation study of the depth-aware module.

| Method | BC ↑ | CS ↓ | Corr ↑ | Inter ↑ |
|---|---|---|---|---|
| Linear | <u>0.7882</u> | 63.2225 | 0.0388 | 0.3965 |
| Gamma | 0.4205 | <u>15.2911</u> | <u>0.9649</u> | <u>1.5315</u> |
| Histogram | **0.8465** | 64.5014 | 0.0141 | 0.3329 |
| Depth-aware | 0.2546 | **0.7881** | **0.9915** | **1.8251** |

**Limitations. (1) Assumption of planar scenes and clean separation.** The current framework is optimized for 2D, front-facing layouts with well-isolated text regions. In scenarios involving occlusion, curved surfaces, or noisy segmentation—such as the obstacle example in Figure 3(c)—text extraction and reintegration can still fail. Future work may explore 3D-aware text alignment and self-correcting segmentation to enhance reliability. **(2) Difficulty handling repeated or ambiguous elements.** Although *CoTextor* performs better in repeated-object scenes (e.g., bags with different characters), it may still confuse glyph style or placement when semantic cues are limited. This highlights the need for more context-aware rewriting mechanisms that leverage both visual and linguistic consistency. Beyond static images, future extensions could incorporate temporal modeling for video and AR content, as well as feedback-driven personalization to support interactive and adaptive co-creation.

**System Capabilities Comparison.** To better understand *CoTextor*'s positioning among visual text editing tools, we compare it with representative baselines along five human-centered capabilities: geometric control (for layout authorship), reversible editing (for iteration), depth-aware composition (for perceptual realism), training-free deployment (for accessibility), and multilingual support (for cultural inclusivity). As shown in Table 2, *CoTextor* is the only system to fulfill all five criteria. This highlights *CoTextor*'s unique value as a modular, user-controllable framework. It enables creators to iteratively adjust layout, appearance, and language without requiring code, annotations, or retraining—features essential for real-world, cross-lingual, and inclusive creative workflows.

Table 2: Comparison of capabilities among visual text editing systems.

| Method | GeoControl | Reversible | Depth | Train-Free | MultiLanguage |
|---|---|---|---|---|---|
| AnyText | ✗ | ✗ | ✗ | ✗ | ✓ |
| TextDiffuser | △ | ✗ | ✗ | ✗ | ✓ |
| ***CoTextor*** | ✓ | ✓ | ✓ | ✓ | ✓ |

Notes: ✓: supported, ✗: not supported, △: partially supported

## 7 Conclusion

We presented *CoTextor*, a modular, training-free framework for controllable multilingual text editing in images. By disentangling the editing process into layered, reversible stages, *CoTextor* supports a user-controllable workflow that centers human input in multilingual and creative contexts. Our depth-aware fusion module enhances photometric and geometric consistency, ensuring perceptual realism in visually complex scenes. Designed for accessibility and inclusivity, *CoTextor* empowers non-technical and culturally diverse users to perform expressive transformations without requiring code, annotations, or retraining. This work contributes to the growing vision of Creative AI systems that foreground transparency, adaptability, and shared authorship.

## Code Available

The code can be found at here.

## Appendix

The appendix can be downloaded from here.

## Acknowledgment

This work was supported by the Open Research Fund of Yunnan Key Laboratory of Quantitative Remote Sensing, and the Talent Training Fund of Kunming University of Science and Technology (Grant No. KKZ3202503073).

## References

[1] Connelly Barnes, Eli Shechtman, Adam Finkelstein, and Dan B Goldman. Patchmatch: A randomized correspondence algorithm for structural image editing. *ACM Trans. Graph.*, 28(3):24, 2009.

[2] Anil Kumar Bhattacharyya. On a measure of divergence between two statistical populations defined by their probability distributions. *Bulletin of the Calcutta Mathematical Society*, 35:99–109, 1943.

[3] Yi Bin, Junrong Liao, Yujuan Ding, Haoxuan Li, Yang Yang, See-Kiong Ng, and Heng Tao Shen. Leveraging weak cross-modal guidance for coherence modelling via iterative learning. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 4630–4639, 2024.

[4] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18392–18402, 2023.

[5] Runze Cai, Nuwan Janaka, Yang Chen, Lucia Wang, Shengdong Zhao, and Can Liu. Pandalens: Towards ai-assisted in-context writing on ohmd during travels. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, pages 1–24, 2024.

[6] Huiwen Chang, Han Zhang, Jarred Barber, AJ Maschinot, Jose Lezama, Lu Jiang, Ming-Hsuan Yang, Kevin Murphy, William T Freeman, Michael Rubinstein, et al. Muse: Text-to-image generation via masked generative transformers. *arXiv preprint arXiv:2301.00704*, 2023.

[7] Haoxing Chen, Zhuoer Xu, Zhangxuan Gu, Yaohui Li, Changhua Meng, Huijia Zhu, Weiqiang Wang, et al. Diffute: Universal text editing diffusion model. *Advances in Neural Information Processing Systems*, 36:63062–63074, 2023.

[8] Jingye Chen, Yupan Huang, Tengchao Lv, Lei Cui, Qifeng Chen, and Furu Wei. Textdiffuser: Diffusion models as text painters. *Advances in Neural Information Processing Systems*, 36, 2024.

[9] Zhangquan Chen, Puhua Jiang, and Ruqi Huang. Dv-matcher: Deformation-based non-rigid point cloud matching guided by pre-trained visual features. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 27264–27274, 2025.

[10] Zhangquan Chen, Chunjiang Liu, and Haobin Duan. A three-phases-lora finetuned hybrid llm integrated with strong prior module in the education context. In *International Conference on Artificial Neural Networks*, pages 235–250. Springer, 2024.

[11] John Joon Young Chung. Artistic user expressions in ai-powered creativity support tools. In *Adjunct Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology*, pages 1–4, 2022.

[12] Joon Gi Chung, Soongi Hong, Junho Choi, and Changhoon Oh. Understanding the dynamics in deploying ai-based content creation support tools in broadcasting systems-benefits, challenges, and directions. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pages 1–28, 2025.

[13] Alloy Das, Sanket Biswas, Prasun Roy, Subhankar Ghosh, Umapada Pal, Michael Blumenstein, Josep Lladós, and Saumik Bhattacharya. Faster: A font-agnostic scene text editing and rendering framework. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1944–1954. IEEE, 2025.

[14] Yadolah Dodge. *The Oxford Dictionary of Statistical Terms*. Oxford University Press, 2003.

[15] Zeyu Dong, Yuyang Yin, Yuqi Li, Eric Li, Hao-Xiang Guo, and Yikai Wang. Panolora: Bridging perspective and panoramic video generation with lora adaptation. *arXiv preprint arXiv:2509.11092*, 2025.

[16] Yue Fu, Michele Newman, Lewis Going, Qiuzi Feng, and Jin Ha Lee. Exploring the collaborative co-creation process with ai: A case study in novice music production. In *Proceedings of the 2025 ACM Designing Interactive Systems Conference*, pages 1298–1312, 2025.

[17] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

[18] Satoshi Iizuka and Edgar Simo-Serra. Deepremaster: temporal source-reference attention networks for comprehensive video enhancement. *ACM Transactions on Graphics (TOG)*, 38(6):1–13, 2019.

[19] AI Jaided. Easyocr: Ready-to-use ocr with 80+ supported languages and all popular image file formats (2021).

[20] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023.

[21] Rick Knops, Bianca Şerban, and Steven Houben. Human-human collaboration enhanced with emerging technologies of ai. In *CHI'23: Workshop on Integrating AI in Human-Human Collaborative Ideation*, 2023.

[22] Praveen Krishnan, Rama Kovvuri, Guan Pang, Boris Vassilev, and Tal Hassner. Textstylebrush: transfer of text aesthetics from a single example. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(7):9122–9134, 2023.

[23] Sanyam Lakhanpal, Shivang Chopra, Vinija Jain, Aman Chadha, and Man Luo. Refining text-to-image generation: Towards accurate training-free glyph-enhanced image generation. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 4372–4381. IEEE, 2025.

[24] Yuqi Li, Kai Li, Xin Yin, Zhifei Yang, Junhao Dong, Zeyu Dong, Chuanguang Yang, Yingli Tian, and Yao Lu. Sepprune: Structured pruning for efficient deep speech separation. *arXiv preprint arXiv:2505.12079*, 2025.

[25] Yuqi Li, Chuangang Yang, Hansheng Zeng, Zeyu Dong, Zhulin An, Yongjun Xu, Yingli Tian, and Hao Wu. Frequency-aligned knowledge distillation for lightweight spatiotemporal forecasting. *arXiv:2507.02939*, 2025.

[26] Yuyuan Li, Chaochao Chen, Yizhao Zhang, Weiming Liu, Lingjuan Lyu, Xiaolin Zheng, Dan Meng, and Jun Wang. Ultrare: Enhancing receraser for recommendation unlearning via error decomposition. *Advances in Neural Information Processing Systems*, 36:12611–12625, 2023.

[27] Yuyuan Li, Chaochao Chen, Xiaolin Zheng, Yizhao Zhang, Zhongxuan Han, Dan Meng, and Jun Wang. Making users indistinguishable: Attribute-wise unlearning in recommender systems. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 984–994, 2023.

[28] Yuyuan Li, Yizhao Zhang, Weiming Liu, Xiaohua Feng, Zhongxuan Han, Chaochao Chen, and Chenggang Yan. Multi-objective unlearning in recommender systems via preference guided pareto exploration. *IEEE Transactions on Services Computing*, 2025.

[29] Bangyan Liao, Zhenjun Zhao, Lu Chen, Haoang Li, Daniel Cremers, and Peidong Liu. Globalpointer: Large-scale plane adjustment with bi-convex relaxation. In *European Conference on Computer Vision*, pages 360–376. Springer, 2024.

[30] Bangyan Liao, Zhenjun Zhao, Haoang Li, Yi Zhou, Yingping Zeng, Hao Li, and Peidong Liu. Convex relaxation for robust vanishing point estimation in manhattan world. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 15823–15832, 2025.

[31] Chang Liu, Haoning Wu, Yujie Zhong, Xiaoyun Zhang, Yanfeng Wang, and Weidi Xie. Intelligent grimm-open-ended visual storytelling via latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6190–6200, 2024.

[32] Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. Repaint: Inpainting using denoising diffusion probabilistic models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11461–11471, 2022.

[33] Bahar Mahmud, Guan Hong, and Bernard Fong. A study of human–ai symbiosis for creative work: Recent developments and future directions in deep learning. *ACM Transactions on Multimedia Computing, Communications and Applications*, 20(2):1–21, 2023.

[34] Thi Tuyet Hai Nguyen, Adam Jatowt, Mickael Coustaty, and Antoine Doucet. Survey of post-ocr processing approaches. *ACM Computing Surveys (CSUR)*, 54(6):1–37, 2021.

[35] Chaojun Ni, Jie Li, Haoyun Li, Hengyu Liu, Xiaofeng Wang, Zheng Zhu, Guosheng Zhao, Boyuan Wang, Chenxin Li, Guan Huang, et al. Wonderfree: Enhancing novel view quality and cross-view consistency for 3d scene exploration. *arXiv preprint arXiv:2506.20590*, 2025.

[36] Chaojun Ni, Xiaofeng Wang, Zheng Zhu, Weijie Wang, Haoyun Li, Guosheng Zhao, Jie Li, Wenkang Qin, Guan Huang, and Wenjun Mei. Wonderturbo: Generating interactive 3d world in 0.72 seconds. *arXiv preprint arXiv:2504.02261*, 2025.

[37] Chaojun Ni, Guosheng Zhao, Xiaofeng Wang, Zheng Zhu, Wenkang Qin, Guan Huang, Chen Liu, Yuyin Chen, Yida Wang, Xueyang Zhang, et al. Recondreamer: Crafting world models for driving scene reconstruction via online restoration. *arXiv preprint arXiv:2411.19548*, 2024.

[38] Chae-Won Park, Vikas Palakonda, Sangseok Yun, Il-Min Kim, and Jae-Mo Kang. Ocr-diff: A two-stage deep learning framework for optical character recognition using diffusion model in industrial internet of things. *IEEE Internet of Things Journal*, 11(15):25997–26000, 2024.

[39] Karl Pearson. Note on regression and inheritance in the case of two parents. *Proceedings of the Royal Society of London*, 58:240–242, 1895.

[40] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022.

[41] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

[42] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 22500–22510, 2023.

[43] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022.

[44] Krishna Srinivasan, Karthik Raman, Jiecao Chen, Michael Bendersky, and Marc Najork. Wit: Wikipedia-based image text dataset for multimodal multilingual machine learning. In *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*, pages 2443–2449, 2021.

[45] Roman Suvorov, Elizaveta Logacheva, Anton Mashikhin, Anastasia Remizova, Arsenii Ashukha, Aleksei Silvestrov, Naejin Kong, Harshith Goka, Kiwoong Park, and Victor Lempitsky. Resolution-robust large mask inpainting with fourier convolutions. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2149–2159, 2022.

[46] Michael J Swain and Dana H Ballard. Color indexing. *International Journal of Computer Vision*, 7(1):11–32, 1991.

[47] Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. Plug-and-play diffusion features for text-driven image-to-image translation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1921–1930, 2023.

[48] Yuxiang Tuo, Wangmeng Xiang, Jun-Yan He, Yifeng Geng, and Xuansong Xie. Anytext: Multilingual visual text generation and editing. In *The Twelfth International Conference on Learning Representations*.

[49] Maciej Wodziński, Marcin Rządeczka, Anastazja Szuła, Marta Sokół, and Marcin Moskalewicz. Visual stereotypes of autism spectrum in dall-e, stable diffusion, sdxl, and midjourney. *arXiv preprint arXiv:2407.16292*, 2024.

[50] Zhuohao Wu, Danwen Ji, Kaiwen Yu, Xianxu Zeng, Dingming Wu, and Mohammad Shidujaman. Ai creativity and the human-ai co-creation model. In *International Conference on Human-Computer Interaction*, pages 171–190. Springer, 2021.

[51] Yi Xin, Qi Qin, Siqi Luo, Kaiwen Zhu, Juncheng Yan, Yan Tai, Jiayi Lei, Yuewen Cao, Keqi Wang, Yibin Wang, et al. Lumina-dimoo: An omni diffusion large language model for multi-modal generation and understanding. *arXiv preprint arXiv: 2510.06308*, 2025.

[52] Yi Xin, Juncheng Yan, Qi Qin, Zhen Li, Dongyang Liu, Shicheng Li, Victor Shea-Jay Huang, Yupeng Zhou, Renrui Zhang, Le Zhuo, et al. Lumina-mgpt 2.0: Stand-alone autoregressive image modeling. *arXiv preprint arXiv:2507.17801*, 2025.

[53] Yukang Yang, Dongnan Gui, Yuhui Yuan, Weicong Liang, Haisong Ding, Han Hu, and Kai Chen. Glyphcontrol: Glyph conditional control for visual text generation. *Advances in Neural Information Processing Systems*, 36, 2024.

[54] Zhongqi Yang, Wenhang Ge, Yuqi Li, Jiaqi Chen, Haoyuan Li, Mengyin An, Fei Kang, Hua Xue, Baixin Xu, Yuyang Yin, et al. Matrix-3d: Omnidirectional explorable 3d world generation. *arXiv preprint arXiv:2508.08086*, 2025.

[55] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. Generative image inpainting with contextual attention. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5505–5514, 2018.

[56] Xinmiao Yu, Xiaocheng Feng, Yun Li, Minghui Liao, Ya-Qi Yu, Xiachong Feng, Weihong Zhong, Ruihan Chen, Mengkang Hu, Jihao Wu, et al. Cross-lingual text-rich visual comprehension: An information theory perspective. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 9680–9688, 2025.

[57] Zhenyu Yu and Chee Seng Chan. Yuan: Yielding unblemished aesthetics through a unified network for visual imperfections removal in generated images. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2025.

[58] Zhenyu Yu, MOHD YAMANI IDNA IDRIS, and Pei Wang. Physics-constrained symbolic regression from imagery. In *2nd AI for Math Workshop@ ICML 2025*, 2025.

[59] Zhenyu Yu, Mohd Yamani Idna Idris, Pei Wang, and Yuelong Xia. Dancetext: Point-driven interactive text and image layer editing using diffusion models. *arXiv preprint arXiv:2504.14108*, 2025.

[60] Zhenyu Yu, Mohd Yamani Idna Idris, Pei Wang, Yuelong Xia, and Yong Xiang. Forgetme: Benchmarking the selective forgetting capabilities of generative models. *Engineering Applications of Artificial Intelligence*, 161:112087, 2025.

[61] Shuang Zeng, Xinyuan Chang, Mengwei Xie, Xinran Liu, Yifan Bai, Zheng Pan, Mu Xu, and Xing Wei. Futuresightdrive: Thinking visually with spatio-temporal cot for autonomous driving. *arXiv preprint arXiv:2505.17685*, 2025.

[62] Shuang Zeng, Dekang Qi, Xinyuan Chang, Feng Xiong, Shichao Xie, Xiaolong Wu, Shiyi Liang, Mu Xu, and Xing Wei. Janusvln: Decoupling semantics and spatiality with dual implicit memory for vision-language navigation. *arXiv preprint arXiv:2509.22548*, 2025.

[63] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3836–3847, 2023.

[64] Yaozong Zheng, Bineng Zhong, Qihua Liang, Guorong Li, Rongrong Ji, and Xianxian Li. Toward unified token learning for vision-language tracking. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(4):2125–2135, 2023.

[65] Yaozong Zheng, Bineng Zhong, Qihua Liang, Ning Li, and Shuxiang Song. Decoupled spatio-temporal consistency learning for self-supervised tracking. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 10635–10643, 2025.

[66] Yaozong Zheng, Bineng Zhong, Qihua Liang, Zhenjun Tang, Rongrong Ji, and Xianxian Li. Leveraging local and global cues for visual tracking via parallel interaction network. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(4):1671–1683, 2022.

[67] Yaozong Zheng, Bineng Zhong, Qihua Liang, Shengping Zhang, Guorong Li, Xianxian Li, and Rongrong Ji. Towards universal modal tracking with online dense temporal token learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

[68] Peihao Zhu, Rameen Abdal, Yipeng Qin, and Peter Wonka. Sean: Image synthesis with semantic region-adaptive normalization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5104–5113, 2020.