Ministry of Higher Education and
Scientific Research
University of Mosul
College of Computer Science and
Mathematics
Department of Computer Science

Text-to-Image Synthesis: Implementing an Intelligent Model for Visualizing Texts

**A Thesis Submitted to the Council of the College of
Computer Science and Mathematics
University of Mosul
as a Partial Fulfillment of Requirements
for the Degree of Doctor of Philosophy in
Computer Science**

By
Haitham Thamer Abdulqader Khalil

Supervised by
Asst. Prof. Dr. Alaa Yaseen Taha Mohammed

2025 A. D. **1447 A. H.**

ABSTRACT

Despite the major developments in generative AI, contemporary text-to-image models struggle to depict extended narratives or abstract, non-visual concepts that necessitate contextual awareness, semantic precision, and temporal coherence. This shortcoming restricts their utilization in storytelling, interactive education, animation, and scientific discourse, areas where visual coherence and interpretability are essential.

This dissertation aims to address this gap by advancing beyond single-prompt drawings to create dependable multimodal methods that convert extensive written sections into coherent images paired with meaningful soundscapes.

This dissertation presents a multimodal dataset of text-image pairs sourced from publicly accessible films and e-books, encompassing a variety of vocabularies, characters, narrative structures, and artistic styles. Additionally, three generative models are presented. First, GANViT, Generative Adversarial Network that integrates Vision-Transformer generators and discriminators to optimize text–image alignment while reducing computational expenses. GANViT achieves comparable perceptual quality with large diffusion models and provides accelerated inference speed. Second, Tale Visualizer is a diffusion-based architecture that conditions each frame on the current caption and the denoising attributes of prior image-text pairs. Third, SoundCrafter is a diffusion-based text-to-sound model that functions within a compact latent space and incorporates semantics via Contrastive Language-Audio Pretraining embeddings. These models aim to achieve the objective of providing a comprehensive system that generates coherent visual narratives and synchronized sound from natural language prompts.

Comprehensive experiments on MS-COCO, AudioCaps, and a custom-collected dataset, which is called Tale Dataset, confirm substantive gains: GANViT reaches an FID of 5.01, Tale Visualizer sustains narrative integrity with CLIP-I 0.75 and FID 33.22, and SoundCrafter converts text to realistic sound at FID 23.45 and IS 7.37, metrics are presented here to underscore the practical impact of the proposed frameworks.

In addition, GANViT and Tale Visualizer transform text into coherent image sequences that function as engaging, cost-effective educational resources, while SoundCrafter promptly generates high-fidelity audio from the same text—collectively enhancing accessibility for learners of all

capabilities, including content creators, educators, and designers, and advancing Sustainable Development Goals (SDG) 4.

This dissertation focuses on advancing multimodal artificial intelligence by integrating image and audio generation, thereby enabling more coherent and inclusive educational techniques.

وزارة التعليم العالي والبحث العلمي
جامعة الموصل
كلية علوم الحاسوب والرياضيات
قسم علوم الحاسوب

توليف نص الى صورة: تنفيذ نموذج ذكي لمرئية النصوص

اطروحة مقدمة
الى مجلس كلية علوم الحاسوب والرياضيات في جامعة الموصل
كجزء من متطلبات نيل شهادة دكتوراه فلسفة في
علوم الحاسوب

من قبل

هيثم ثامر عبدالقادر خليل

بإشراف
أ. م. د. الاء ياسين طه محمد

١٤٤٧ هـ
٢٠٢٥ م

المستخلص

على الرغم من التطورات الكبيرة في الذكاء الاصطناعي التوليدي، تواجه نماذج تحويل (النص إلى صورة) المعاصرة صعوبة في تصوير السرديات المطولة أو المفاهيم المجردة غير المرئية التي تتطلب وعيًا سياقيًا ودقة دلالية وتماسكًا زمنيًا. يحد هذا القصور من استخدامها في سرد القصص والتعليم التفاعلي والرسوم المتحركة والخطاب العلمي، وهي مجالات تُعد فيها التماسك البصري وقابلية التفسير أمرًا بالغ الأهمية. تهدف هذه الأطروحة إلى سدّ هذه الفجوة من خلال تجاوز الرسومات أحادية التوجيه إلى ابتكار أساليب متعددة الوسائط موثوقة تُحوّل المقاطع المكتوبة المطولة إلى صور متماسكة مقترنة بمشاهد صوتية مؤثرة.

اقترحت هذه الأطروحة مجموعة بيانات شاملة متعددة الوسائط لأزواج النصوص والصور، مستمدة من أفلام وكتب إلكترونية متاحة للعامة، وتشمل مجموعة متنوعة من المفردات والشخصيات والهياكل السردية والأساليب الفنية. بالإضافة إلى ذلك، تقدم ثلاثة نماذج توليدية. أولًا، GANViT، وهو نموذج GAN مُبسّط يدمج مُولّدات ومُميّزات Vision Transformer لتحسين محاذاة النص والصورة مع تقليل التكاليف الحسابية. يحقق GANViT جودة إدراكية تُضاهي نماذج الانتشار الكبيرة، وتوفير سرعة استدلال مُسرّعة. ثانيًا، يُعد Tale Visualizer بنيةً قائمةً على الانتشار تُحدد كل إطار بناءً على التسمية التوضيحية الحالية وخصائص إزالة الضوضاء لأزواج الصور والنصوص السابقة. ثالثًا، يُعد SoundCrafter نموذجًا قائمًا على الانتشار لتحويل النص إلى صوت، ويعمل ضمن مساحة كامنة مُدمجة، ويُدمج الدلالات عبر تضمينات CLAP.

تهدف هذه النماذج إلى تحقيق هدف توفير نظام شامل يُولّد سرديات بصرية متماسكة وصوتًا مُتزامنًا من مُحفزات اللغة الطبيعية. تؤكد التجارب الشاملة على MS-COCO وAudioCaps ومجموعة بيانات تم جمعها خصيصًا، والتي تسمى Tale Dataset، على مكاسب جوهرية: يصل GANViT إلى FID 5.01، ويحافظ Tale Visualizer على سلامة السرد مع CLIP-I 0.75 وFID 33.22، ويحول SoundCrafter النص إلى صوت بيئي محيطي عند FID 23.45 و IS 7.37، وهي المقاييس المقدمة هنا للتأكيد على التأثير العملي للنماذج المقترحة.

بالإضافة إلى ذلك، يقوم GANViT وTale Visualizer بتحويل النص إلى تسلسلات رسومية متماسكة تعمل كموارد تعليمية جذابة وفعالة من حيث التكلفة، بينما يقوم SoundCrafter على الفور بإنشاء صوت عالي الدقة من نفس النص – مما يعزز بشكل جماعي إمكانية الوصول للطلاب وتعزيز أهداف التنمية المستدامة (الهدف الرابع).