Text to Video Creation using Decomposition Model

Ms. Shaheen Mujawar[1], Mr. Aman B[2], Mr. Athar A[3], Ms. Mrunal M[4], Ms. Saniya B[5]

[1]Department of Computer Science and Engineering, S. G. Balekundri Institute of Technology, Belagavi,Karnataka, India (Assistant Professor), shaheenm@sgbit.edu.in

[2]Department of Computer Science and Engineering, S. G. Balekundri Institute of Technology, Belagavi,Karnataka, India (Student), amanbelgaumi6@gmail.com

[3]Department of Computer Science and Engineering, S. G. Balekundri Institute of Technology, Belagavi,Karnataka, India (Student), ansariathar855@gmail.com

[4]Department of Computer Science and Engineering, S. G. Balekundri Institute of Technology, Belagavi,Karnataka, India (Student), mrunalmutagekar@gmail.com

[5]Department of Computer Science and Engineering, S. G. Balekundri Institute of Technology, Belagavi,Karnataka, India (Student), saniyabagwan13456@gmail.com

Abstract

In this paper, "Scripted Vision: An Exploration of Text-to-Video Dynamics," we introduce a system aimed at generating videos from textual prompts using Diffusion Models. This system offers advanced features such as seed randomization, frame count adjustment, and inference step configuration, all accelerated by Graphical Processing Unit (GPU) technology for efficient processing. Our research demonstrates that the generated videos exhibit diverse content representations corresponding to the provided prompts, and the system's adaptability extends to platforms like Google Colab via Gradio for real-time interaction and video generation with prompt input and feedback. Leveraging training data from LAION5B, ImageNet, WebVid, and other public datasets, our system showcases the fusion of natural language understanding and computer vision techniques, empowering users to translate textual descriptions into compelling visual narratives. Its adaptability and user-friendly interface offer broader applications in multimedia content creation, education, entertainment industries, and immersive storytelling experiences, bridging the gap between text-based prompts and video content creation.

Keywords: Diffusion Model (DM), Diffusion Pipeline (DP), Natural Language Processing (NLP), Variational Autoencoder (VAE)

I. INTRODUCTION

Text-to-video generation stands as a cutting-edge application within the deep learning domain, seamlessly translating written text into corresponding video content. The introduction of the innovative method, "Scripted Vision: An Exploration of Text-to-Video Dynamics" in the deep learning landscape represents a significant leap forward. Scripted Vision redefines the boundaries of content generation, offering a novel approach that marries the realms of natural language understanding and computer vision. By leveraging state-of the-art Diffusion Models, Scripted Vision empowers users to transform textual prompts into vivid and coherent video sequences with remarkable fidelity. At its core, Scripted Vision harnesses the power of Diffusion Models, the diffusion model serves as a fundamental component in translating textual prompts into coherent and visually compelling video sequences. The process initiates with encoding the textual prompt into a latent representation, which serves as the starting point for video generation. Through a series of iterative diffusion steps, the model refines this latent representation, guided by the information embedded within the text [1]. Throughout the diffusion process, contextual cues from the textual input are incorporated to guide the generation of video content, ensuring relevance and coherence. Crucially, the model maintains temporal coherence by considering dependencies between consecutive video frames, resulting in smooth transitions and continuity. As the diffusion progresses, the model continuously enhances the quality and realism of generated video frames through a combination of sampling strategies and probabilistic modelling approaches. Once convergence is reached, the final latent representation is decoded into a coherent video sequence, reflecting the essence of the given textual prompt. In essence, the diffusion model serves as a bridge between textual inputs and video outputs, enabling the synthesis of high-quality and contextually relevant video content from textual prompts in text-to-video generation systems [2].

The utilization of Diffusion Pipeline and DPMSolverMultistepScheduler algorithms, coupled with techniques like VAE slicing, further enhances the efficiency and effectiveness of the generative process. Here, The Diffusion Pipeline forms the backbone of the text-to-video generation process [3]. It comprises a series of interconnected modules designed to preprocess, transform, and refine the input textual prompts into visually compelling video content. Within the pipeline, various stages such as text embedding, feature extraction, and content synthesis are seamlessly integrated to facilitate the smooth transition from text to video. The utilization of a well-structured pipeline ensures the systematic flow of data and operations, optimizing resource utilization and enhancing the overall efficiency of the generative process. And the DPMSolverMultistepScheduler algorithm incorporates a suite of techniques to manage the diffusion process in text-to-video generation systems efficiently. It utilizes gradient descent optimization algorithms like stochastic gradient descent (SGD), Adam, or RMSprop to iteratively update model parameters based on the loss function gradients, thereby enhancing the quality of generated content. Learning rate scheduling strategies such as exponential decay or cosine annealing dynamically adjust the learning rate during training to improve convergence and stability [6]. Annealing techniques like linear or exponential annealing gradually modify the diffusion step size or temperature parameter to guide the diffusion towards high-quality solutions. Adaptive step size adjustment algorithms like adaptive Metropolis-Hastings (AMH) or adaptive Langevin dynamics dynamically adapt the diffusion step size based on convergence behaviour to improve sampling efficiency. Multi-step sampling techniques such as Metropolis-adjusted Langevin

algorithm (MALA) or Hamiltonian Monte Carlo (HMC) accelerate convergence and improve sample quality by performing multiple diffusion steps per iteration [7]. Convergence monitoring algorithms track diffusion progress and determine convergence criteria, ensuring the diffusion process converges to high-quality solutions. Additionally, parallelization and distributed computing techniques optimize computational resource utilization, accelerating the diffusion process through data or model parallelism across multiple devices or nodes. These combined algorithms and techniques empower the DPMSolverMultistepScheduler to effectively orchestrate the diffusion process and generate high-quality video content from textual prompts efficiently [8].

Variational Autoencoder (VAE) slicing is a sophisticated technique employed within the diffusion model framework to enhance the diversity and richness of generated video content. By manipulating the latent space of the VAE, the slicing technique enables the generation of multiple variations of a given textual prompt. This process involves exploring different regions of the latent space and sampling latent vectors to produce diverse outputs. Through VAE slicing, Scripted Vision can offer users a wide array of video interpretations for a single textual input, fostering creativity and exploration in content generation [9].

GPU acceleration plays a pivotal role in enabling the computational demands of Scripted Vision. By leveraging the parallel processing capabilities of GPUs, the system achieves unprecedented speeds in video generation, facilitating real-time feedback and interaction through a user-friendly web interface. This computational prowess ensures that users can effortlessly input textual prompts and witness the generation of corresponding video content in near real-time, enhancing user engagement and accessibility. The effectiveness of the diffusion model in transforming raw textual data into coherent video narratives relies heavily on the quality and diversity of the trained datasets. LIaon5b, ImageNet, and WebVid are among the meticulously curated datasets used to train the diffusion model. LIaon5b provides a rich source of diverse textual prompts, ranging from everyday language to specialized domains, enabling the model to capture a wide spectrum of semantic concepts [11]. ImageNet, a benchmark dataset in computer vision, offers a vast collection of labeled images, enriching the visual understanding capabilities of the model. Additionally, WebVid dataset encompasses a diverse range of video content sourced from the web, facilitating the synthesis of realistic and contextually relevant video sequences. By leveraging these diverse datasets, Scripted Vision ensures the fidelity and coherence of generated video content across different domains and scenarios [13].

Moreover, Scripted Vision offers advanced customization options, allowing users to adjust parameters such as seed randomization, frame count, and inference steps to tailor the generated videos to their preferences. This flexibility ensures that users can fine-tune the output to suit their specific needs, whether they require short, concise videos or longer, more elaborate sequences. In essence, Scripted Vision represents a convergence of cutting-edge technologies in natural language processing, computer vision, and GPU acceleration. By pushing the boundaries of text-to-video generation, Scripted Vision opens up new avenues for creative expression and content generation, empowering users to transform written text into captivating visual narratives with unprecedented ease and efficiency.

II. LITERATURE SURVEY

A literature survey is a critical component of any research project. It involves reviewing and analyzing existing literature, research papers, and other relevant sources to gain a comprehensive understanding of the current state of knowledge on a particular topic.

Table 1: Literature Survey

S.No	Title/Year/Authors	Methodology Followed	Observation
[1]	Make-A-Video: Text-To-Video Generation Without Text-Video Data. [2023] Uriel Singer, Adam Polyak	Pre-trained text-to-image (T2I) model and Super-resolution models.	The paper presents Make-A-Video, a method for generating videos from text without text-video data. It uses text-to-image models and unsupervised video models to create realistic and diverse videos. It outperforms existing methods and introduces a new dataset and applications for text-to-video generation. The paper is clear, thorough, and impactful [1].
[2]	Newsgist: video generation from news stories.[2023] M. S. Karthika Devi & R. Baskaran	Named Entity Recognition (NER), Convolutional neural network (CNN), Logistic Regression, Face swapping	The authors introduce a novel video generation system from news stories, utilizing collaborative learning and knowledge representation. While achieving high accuracy in dialogue extraction and integration into videos across domains and languages, the paper overlooks key limitations. Future work should explore video quality, system scalability, ethical considerations, and evaluation methods for enhanced system applicability and performance [2].
[3]	Tune-A-Video: One-Shot Tuning of Image Diffusion Models for Text-to-Video Generation [2023] Jay Zhangjie Wu, Yixiao Ge, Xintao Wang, Stan Weixian Lei, Yuchao Gu, Yufei Shi1 Wynne Hsu, Ying Shan, Xiaohu Qie, Mike Zheng Shou.	Denoising diffusion probabilistic models (DDPMs), Fine-Tuning and Network Inflation.	Models for Text-to-Video Generation, have made commendable progress in the realm of text-to-video generation. However, it is essential to acknowledge certain drawbacks and limitations inherent in their approach. While the one-shot tuning of image diffusion models is a novel and efficient strategy, it may encounter challenges in capturing the full spectrum of complexities present in diverse textual inputs. Furthermore, the model's performance might vary in scenarios that require nuanced adjustments beyond the capabilities of one-shot tuning [3].

[4]	Align your Latents: High-Resolution Video Synthesis with Latent Diffusion Models 2023 Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, Karsten Kreis.	Latent diffusion models (LDMs).	"Align your Latents: High-Resolution Video Synthesis with Latent Diffusion Models" presents an intriguing approach to high-resolution video synthesis. However One potential concern could be the computational complexity associated with the proposed latent diffusion models. High-resolution video synthesis often demands significant computational resources, which may hinder its practicality for real-time applications or for researchers with limited access to powerful hardware. Additionally, the effectiveness of the model might be highly dependent on the quality and quantity of training data, raising questions about its robustness across diverse scenes and scenarios [4].
[5]	VideoComposer: Compositional Video Synthesis with Motion Controllability 2023 Xiang Wang, Hangjie Yuan, Shiwei Zhang, Dayou Chen, Jiuniu Wang, Yingya Zhang, Yujun Shen, Deli Zhao, Jingren Zhou	Decomposition of Video Conditions, Latent Diffusion Model Training, Incorporation of Temporal Guidance, Introduction of STC-encoder, Efficient Utilization of Control Signals, Demonstration of Efficacy.	Video Composer enhances control in visual generative models for video synthesis by decomposing videos into textual, spatial, and temporal conditions. It employs a latent diffusion model and a video-specific motion vector, with a unified spatial-temporal consistency (STC)-encoder ensuring cross-frame attention mechanisms. This approach allows flexible video composition while maintaining synthesis quality. Video Composer demonstrates effectiveness through creative generative tasks, overcoming challenges in intricate temporal video structures [5].
[6]	Latent Video Diffusion Models for High-Fidelity Long Video Generation 2023 Yingqing He, Tianyu Yang, Yong Zhang, Ying Shan, Qifeng Chen, The Hong Kong University of Science and Technology	Problem Identification, Selection of Generative Models, Development of LVDM, Hierarchical Framework, Performance Mitigation Techniques, Benchmark Evaluation, Extension to Text-to-Video Generation, Public Availability.	The paper addresses challenges in generating photorealistic videos. LVDM, an efficient video diffusion model, achieves state-of-the-art results in short and long video generation. The hierarchical framework extends videos beyond training length, and conditional latent perturbation mitigates performance degradation. LVDM excels in both short and long video generation, showcasing effectiveness in text-to-video synthesis. Contributions include LVDM as a diffusion-based video synthesis baseline, with publicly available code and pre-trained models [6].

[7]	A Systematic Literature Review on Text Generation Using Deep Neural Network Models 2022 Noureen Fatima, Ali Shariq Imran, (member, ieee), Zenun Kastrati, Sher Muhammad Daudpota	PRISMA framework (Preferred Reporting Items for Systematic Reviews and Meta-Analyses)	The authors conducted a thorough literature review on text generation using deep neural network models, following the PRISMA framework. The review covers model types, quality metrics, challenges, limitations, and future trends. However, improvement is needed in detailing the data extraction process, including the number of studies included/excluded, distribution across domains/languages, and statistical methods for data comparison. Enhancing these details would boost transparency and reproducibility [7].
[8]	Make It Move: Controllable Image-to-Video Generation with Text Descriptions 2020 Yaosi Hu1, Chong Luo2, Zhenzhong Chen1	Text-Image-to-Video (TI2V) task based on MNIST and CATER , Motion Anchor-based video Generator (MAGE) Model.	In this paper, the authors propose a novel task of text-image-to-video generation (TI2V), which aims to generate videos from a static image and a text description that specifies the desired appearance and motion of the video. To solve this challenging task, they introduce a motion anchor-based video generator (MAGE), which can align appearance and motion from different modalities and model the uncertainty and diversity of text descriptions. They evaluate their model on two new video-text paired datasets based on MNIST and CATER, and demonstrate its effectiveness and potential over several baselines. The paper is well written, novel, and valuable for the field of text generation and computer vision [8].

III. RESEARCH METHODS

The methodology employed in the "Scripted Vision: An Exploration of Text-to-Video Dynamics", revolves around seamlessly integrating cutting-edge technologies from natural language processing, computer vision, and GPU acceleration to facilitate the generation of video content from textual prompts. At its core, this methodology harnesses Diffusion Models, a class of generative AI models renowned for their ability to produce realistic images and videos. These models undergo iterative diffusion processes, continuously refining and enhancing generated content to ensure fidelity to the provided textual prompts. To enable efficient video generation, the project utilizes the Diffusion Pipeline alongside the DPMSolverMultistepScheduler algorithms. These algorithms, combined with techniques such as VAE slicing, significantly enhance the effectiveness of the generative process. Furthermore, GPU acceleration plays a pivotal role in meeting the computational demands of Scripted Vision. By leveraging the parallel processing capabilities of GPUs, the system achieves unprecedented speeds in video generation, facilitating near real-time feedback and interaction through an intuitive web interface. A crucial aspect of the methodology is providing advanced customization options to users. This encompasses parameters like seed randomization, frame count, and inference steps, empowering users to tailor generated videos to their specific preferences. Such flexibility ensures that users can adjust the output according to their requirements, whether they seek concise videos or more elaborate sequences. In practice, the methodology involves several steps. Initially, the user inputs a textual prompt via the provided interface. Subsequently, the system employs Diffusion Models to generate corresponding video content, taking into account provided parameters such as seed value and frame count to customize the output. Once generated, the resulting video is presented to the user through the interface, enabling real-time interaction and feedback. Moreover, the project incorporates mechanisms for caching examples, particularly beneficial when running the system on GPU-enabled devices. This caching mechanism enhances performance by storing and reusing generated examples, thereby reducing computational overhead. Overall, the methodology of "Scripted Vision: An Exploration of Text-to-Video Dynamics" represents a convergence of state-of-the-art techniques in deep learning. It enables the transformation of written text into compelling visual narratives with unprecedented ease and efficiency. Through a combination of advanced algorithms, GPU acceleration, and user-centric design, Scripted Vision redefines the landscape of text-to-video generation, opening up new avenues for creative expression and content creation.

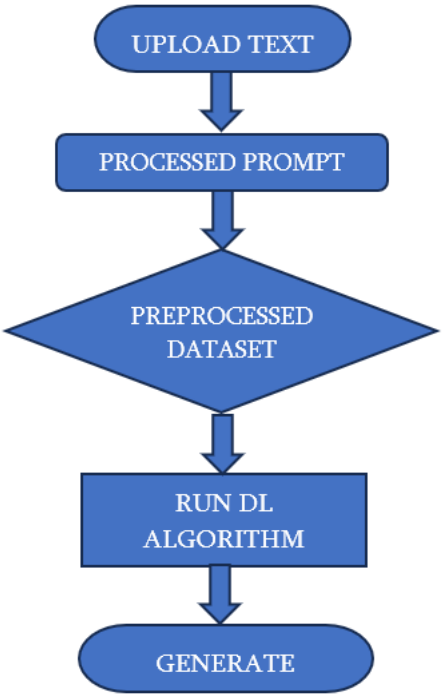

Fig 3.2: Functional Flowchart

The systematic journey from raw text data to the creation of a final video, as depicted in Figure 1, involves several key stages. "Scripted Vision: An Exploration of Text-to-Video Dynamics " streamlines text-to-video creation via a user-friendly web interface. Beginning with Text Prompt Processing, user-provided text undergoes preprocessing for Video Generation. Advanced techniques like Diffusion Models and VAE slicing shape videos according to user parameters. The User Interface Management presents generated videos for real-time interaction, managing caching for performance. Data flows from User to Processing, then Generation, with user preferences directly influencing video creation. Generated videos are delivered back to the user. Optional Cached Examples store previous videos for

improved system performance. This interconnected process efficiently transforms textual prompts into engaging visual narratives, fostering user creativity and engagement.

IV. APPLICATIONS

The following are some of the most prevalent and commonly utilized uses of Scripted Vision: An Exploration of Text-to-Video Dynamics:

A. Personalized Video Content:
By analyzing the emotions expressed in textual input, the system tailors video content to resonate with specific emotional states, enhancing user engagement and creating a more personalized viewing experience.

B. Educational Content Enhancement:
In educational videos, the system can detect and reflect the emotions conveyed in textual content, optimizing the selection of visual aids, expressions, and scenarios to create a more emotionally resonant and effective learning experience.

C. Interactive Storytelling:
Enabling the emotion recognition system in text-to-video projects opens avenues for interactive storytelling where the emotional nuances of the narrative dynamically shape the visual and auditory elements of the video, creating a more immersive and engaging experience.

D. Marketing and Advertisement:
For marketing videos, the system ensures that the emotional undertones of the marketing message align with the visual content, maximizing the impact on the audience and creating more compelling and emotionally resonant advertisements.

E. Human-Computer Interaction in AI Chatbots:
The system's emotional analysis capabilities can elevate AI chatbots' responses by adjusting their tone and expressions based on users' emotions, fostering more natural and relatable interactions.

F. Accessibility Features for Differently-Abled Users:
Inclusive technology applications can utilize the emotion recognition system to enhance accessibility features. For example, adapting video content based on emotional cues can provide a more tailored experience for users with varying emotional needs.

V. RESULT AND DISCUSSION

The figure 1 depicts the process of converting a text description into a series of images. Starting with the text input as “Tiger in jungle”, the progression moves from left to right, with each image becoming increasingly distorted and abstract. These images represent steps in the intricate process of translating textual information into visual content. This visual representation could be intriguing for a Scripted Vision project, showcasing how AI algorithms generate images based on textual input.

Fig 1: Snapshot of result for text input as “Tiger in jungle”

The figure 2 depicts the process of converting a text description into a series of images. Starting with the sentence “Honey bee collecting pollen on a blooming sunflower,” the progression moves from left to right, with each image becoming increasingly detailed and dynamic. These images represent steps in the intricate process of translating textual information into visual content for a Scripted Vision project. The transformation showcases how AI algorithms generate a sequence of images based on descriptive sentences, effectively bridging the gap between text and video.

Fig 2: Snapshot of result for text input as "Honey bee collecting pollen on blooming sunflower"

VI. CONCLUSION

In conclusion, Scripted Vision revolutionizes text-to-video generation by translating written text into captivating visual narratives. By leveraging advanced techniques like Diffusion Models and Variation Auto encoder slicing, it offers users unprecedented creative freedom and efficiency. With GPU acceleration and meticulously curate datasets, Scripted Vision ensures high-quality, diverse video outputs across various domains. Its advanced customization options further enhance user experience, making it a powerful tool for content creation. In essence, Scripted Vision represents a breakthrough in merging natural language processing and computer vision, empowering users to effortlessly transform text into videos, in a new creative expression.

VII. REFERENCES

[1]. Singer, U., Polyak, A., Hayes, T., Yin, X., An, J., Zhang, S., Hu, Q., Yang, H., Ashual, O., Gafni, O., Parikh, D., Gupta, S., Taigman, Y., et al. (2023). Make-A-Video: Text-to-Video Generation without Text-Video Data. In Proceedings of the International Conference on Learning Representations (ICLR).

[2] Devi, M. S. K., & Baskaran, R. (2023). Newsgist: Video Generation from News Stories. Automatika, 64(4), 1026–1037. DOI: 10.1080/00051144.2023.2241774

[3] Wu, J. Z., Ge, Y., Wang, X., Lei, S. W., Gu, Y., Shi, Y., Hsu, W., Shan, Y., Qie, X., Shou, M. Z., et al. (2023). Tune-A-Video: One-Shot Tuning of Image Diffusion Models for Text-to-Video Generation. In Proceedings of the International Conference on Learning Representations (ICLR).

[4] Blattmann, A., Rombach, R., Ling, H., Dockhorn, T., Kim, S. W., Fidler, S., & Kreis, K. (2023). Align your Latents: High-Resolution Video Synthesis with Latent Diffusion Models. arXiv:2304.08818v2 [cs.CV].

[5] Wang, X., Yuan, H., Zhang, S., Chen, D., Wang, J., Zhang, Y., Shen, Y., Zhao, D., & Zhou, J. (2023). VideoComposer: Compositional Video Synthesis with Motion Controllability. arXiv:2306.02018v2 [cs.CV].

[6] He, Y., Yang, T., Zhang, Y., Shan, Y., & Chen, Q. (2023). Latent Video Diffusion Models for High-Fidelity Long Video Generation. arXiv:2211.13221v2 [cs.CV].

[7] Fatima, N., Imran, A. S., Kastrati, Z., Daudpota, S. M., Soomro, A., & Duan, A. S. (2022). A Systematic Literature Review on Text Generation Using Deep Neural Network Models. IEEE Access, 10, 1–14. DOI: 10.1109/ACCESS.2022.3174108

[8] Hu, Y., Luo, C., & Chen, Z. (2022). Make It Move: Controllable Image-to-Video Generation with Text Descriptions. In 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).

[9] Hayes, T., Zhang, S., Yin, X., Pang, G., Sheng, S., Yang, H., Ge, S., Hu, Q., & Parikh, D. (2022). MUGEN: A Playground for Video-Audio-Text Multimodal Understanding and Generation.

[10] Hong, W., Ding, M., Zheng, W., Liu, X., Tang, J., et al. (2022). CogVideo: Large-scale Pretraining for Text-to-Video Generation via Transformers. arXiv preprint arXiv:2205.15868v1 [cs.CV].

[11] Wu, C., Huang, L., Zhang, Q., Li, B., Ji, L., Yang, F., Sapiro, G., & Duan, N. (2021). GODIVA: Generating Open-DomaIn Videos from nAtural Descriptions. arXiv preprint arXiv:2104.14806v1 [cs.CV].

[12] Wu, C., Liang, J., Ji, L., Yang, F., Fang, Y.,

Jiang, D., & Duan, N. (2021). NU¨ WA: Visual Synthesis Pre-training for Neural visUal World creAtion. arXiv preprint arXiv:2111.12417v1 [cs.CV].

[13] Iqbal, T., Qureshi, S., et al. (2020). The survey: Text generation models in deep learning. Journal of King Saud University - Computer and Information Sciences, 34(6), 2515–2528.

[14] Kim, D., Joo, D., Kim, J., et al. (2020). TiVGAN: Text to Image to Video Generation With Step-by-Step Evolutionary Generator. IEEE Access, 8, 153113–153122.

[15] Li, Y., Min, M. R., Shen, D., Carlson, D., Carin, L., et al. (2018). Video Generation from Text.

[16] Zhang, Y., Tsipidi, E., Schriber, S., Kapadia, M., Gross, M., & Modi, A. (2019). Generating Animations from Screenplays.

[17] Wang, Z., Dai, M., & Lundgaard, K. (2023). Text-to-Video: a Two-stage Framework for Zero-shot Identity-agnostic Talking-head Generation. [cs.CV].

[18] Yao, L., Torabi, A., Cho, K., Ballas, N., Pal, C., Larochelle, H., & Courville, A. (2011). Describing Videos by Exploiting Temporal Structure. Pattern Recognition, 44, 2588–25