

การออกแบบและพัฒนาวิธีการป้องกันภาพที่มีลิขสิทธิ์ถูกใช้ใน Generative AI

ภาควัต หาดสมบัติ

TNI

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต

บัณฑิตศึกษา สาขาเทคโนโลยีสารสนเทศ

สถาบันเทคโนโลยีไทย-ญี่ปุ่น

ปีการศึกษา 2568

DESIGN AND DEVELOPMENT OF METHODS FOR PROTECTING COPYRIGHTED IMAGES FROM
UNAUTHORIZED USE IN GENERATIVE AI SYSTEM

Pakawat Hadsombut

TNI

A Thesis Submitted in Partial Fulfillment of the Requirements
for the Degree of Master of Science Program in Information Technology
Graduate Studies

Thai-Nichi Institute of Technology

Academic Year 2025

หัวข้อวิทยานิพนธ์

การออกแบบและพัฒนาวิธีการป้องกันภาพที่มีลิขสิทธิ์ถูกใช้
ใน Generative AI

โดย

ภควัต หาดสมบัติ

สาขาวิชา

เทคโนโลยีสารสนเทศ

อาจารย์ที่ปรึกษาวิทยานิพนธ์

ดร.สะพรั่งสิทธิ์ มฤตสาร

บัณฑิตศึกษา สถาบันเทคโนโลยีไทย-ญี่ปุ่น อนุมัติให้บัณฑิตวิทยานิพนธ์ฉบับนี้เป็นส่วนหนึ่ง
ของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรบัณฑิต

.....รองอธิการบดีฝ่ายวิชาการ
(รองศาสตราจารย์ ดร.วรากร ศรีเชวงทรัพย์)
วันที่.....เดือน.....พ.ศ.....

คณะกรรมการสอบวิทยานิพนธ์

.....ประธานกรรมการ
(รองศาสตราจารย์ ดร.ทิมพันธุ์ เจริญพงษ์)

.....กรรมการ
(ดร.ประมุข บุญเสียง)

.....กรรมการ
(ว่าที่ร้อยตรี ดร.พิชิตชัย คำอินทร์)

.....อาจารย์ที่ปรึกษาวิทยานิพนธ์
(ดร.สะพรั่งสิทธิ์ มฤตสาร)

ภควัต หาดสมบัติ : การออกแบบและพัฒนาวิธีการป้องกันภาพที่มีลิขสิทธิ์ถูกใช้ใน Generative AI. อาจารย์ที่ปรึกษา : ดร.สพรั่งสิทธิ์ มฤตสาร, 104 หน้า.

ในปัจจุบัน โมเดล Generative AI ที่อาศัย Diffusion ได้ถูกนำมาใช้อย่างแพร่หลายในการสร้างและดัดแปลงภาพ ส่งผลให้เกิดความกังวลเกี่ยวกับการคุ้มครองลิขสิทธิ์และการนำเนื้อหาต้นฉบับไปใช้ซ้ำโดยไม่ได้รับอนุญาต แม้ว่าวิธีการป้องกันส่วนใหญ่จะมุ่งเน้นไปที่การป้องกันการนำภาพในขั้นตอนการฝึกหรือปรับแต่งโมเดล แต่การป้องกันในกระบวนการ Image-to-Image (img2img) ยังได้รับการศึกษาอย่างจำกัด งานวิจัยนี้นำเสนอวิธีการป้องกันภาพด้วยการฝัง Perturbation ใน Latent Space สำหรับ Stable Diffusion 1.5 (โหมด img2img) โดยทำการฝัง Perturbation ที่มีโครงสร้างผ่านกระบวนการปรับเหมาะแบบวนซ้ำด้วย Gradient และฟังก์ชัน Loss หลายองค์ประกอบ วิธีการนี้มีเป้าหมายเพื่อรักษาความสอดคล้องเชิงการรับรู้และเชิงความหมายกับภาพต้นฉบับ ขณะเดียวกันส่งผลต่อกระบวนการ diffusion เพื่อให้ผลลัพธ์ที่สร้างขึ้นเกิดความแตกต่างเมื่อมีการนำภาพที่ผ่านการป้องกันไปใช้ซ้ำภายใต้เงื่อนไขเดียวกัน ผลการทดลองจากภาพจำนวน 20 ภาพใน 4 หมวดหมู่แสดงให้เห็นว่า ภาพที่ผ่านการป้องกันยังคงมีความคล้ายคลึงกับภาพต้นฉบับในระดับสูง (SSIM = 0.7748, PSNR = 26.59 dB, CLIP = 0.9553, LPIPS = 0.2031) ซึ่งบ่งชี้ถึงการสูญเสียคุณภาพในระดับต่ำ อย่างไรก็ตาม เมื่อเข้าสู่กระบวนการ img2img พบว่าผลลัพธ์มีความแตกต่างอย่างชัดเจนเมื่อเทียบกับภาพต้นฉบับ (SSIM = 0.5499, PSNR = 17.14 dB, CLIP = 0.8153, LPIPS = 0.4760) สะท้อนถึงความแตกต่างในเชิงโครงสร้าง การรับรู้ และความหมาย ผลลัพธ์ดังกล่าวแสดงให้เห็นว่า Perturbation ใน Latent Space สามารถเป็นกลไกที่มีประสิทธิภาพและควบคุมได้สำหรับการป้องกันลิขสิทธิ์ใน Stable Diffusion 1.5.

บัณฑิตศึกษา

สาขาวิชา เทคโนโลยีสารสนเทศ

ปีการศึกษา 2568

ลายมือชื่อนักศึกษา

ลายมือชื่ออาจารย์ที่ปรึกษา

PAKAWAT HADSOMBUT : DESIGN AND DEVELOPMENT OF METHODS FOR PROTECTING
COPYRIGHTED IMAGES FROM UNAUTHORIZED USE IN GENERATIVE AI SYSTEM.
ADVISOR : DR.SAPRANGSIT MRUETUSATORN, 104 PP.

Diffusion-based Generative AI models have become widely used for image synthesis and manipulation, raising concerns about copyright protection and unauthorized reuse of original content. While most existing approaches focus on preventing the use of copyrighted images during model training or fine-tuning, protection in image-to-image (img2img) generation remains relatively underexplored. This study proposes a latent-space perturbation method for protecting copyrighted images in Stable Diffusion 1.5 (img2img mode), where structured perturbations are embedded into the latent representation through an iterative gradient-based optimization process guided by a multi-component loss function. The method aims to preserve perceptual and semantic consistency with the original image while influencing the downstream diffusion process so that generated outputs deviate when protected images are reused under identical conditions. Experimental results on 20 images across four categories show that protected images maintain high similarity to their originals (SSIM = 0.7748, PSNR = 26.59 dB, CLIP = 0.9553, LPIPS = 0.2031), indicating minimal perceptual degradation. However, when processed through the img2img pipeline, outputs generated from protected images exhibit substantial divergence compared to those from original images (SSIM = 0.5499, PSNR = 17.14 dB, CLIP = 0.8153, LPIPS = 0.4760), reflecting clear structural, perceptual, and semantic variation. These findings suggest that latent-space perturbation can serve as an effective and controlled mechanism for copyright protection in Stable Diffusion 1.5.

Graduate Studies

Student's Signature.....

Field of Study Information Technology

Advisor's Signature.....

Academic Year 2025

กิตติกรรมประกาศ

วิทยานิพนธ์ฉบับนี้สามารถสำเร็จลุล่วงได้ด้วยความรู้ความกรุณาและการสนับสนุนอย่างดียิ่งจากอาจารย์ที่ปรึกษา ดร.สพรั่งสิทธิ์ มฤตุสาธร์ ซึ่งได้กรุณาถ่ายทอดความรู้ ให้คำแนะนำที่เป็นประโยชน์ ตลอดจนให้ความช่วยเหลือในทุกขั้นตอนของการดำเนินงาน รวมทั้งตรวจสอบและชี้แนะแนวทางในการปรับปรุงแก้ไขข้อบกพร่องต่างๆ อย่างต่อเนื่อง ผู้วิจัยขอกราบขอบพระคุณเป็นอย่างสูงไว้ ณ โอกาสนี้

ผู้วิจัยขอขอบพระคุณคณะกรรมการสอบวิทยานิพนธ์ ได้แก่ รองศาสตราจารย์ ดร.ทิมพันธ์ เจริญพงษ์ ดร.ประมุข บุญเสียง และว่าที่ร้อยตรี ดร.พิชิตชัย คำอินทร์ ผู้ให้คำแนะนำข้อแก้ไขที่ควรปรับปรุงจนจบวิทยานิพนธ์ฉบับนี้สมบูรณ์

ขอขอบพระคุณเพื่อนร่วมศึกษาระดับบัณฑิตศึกษาทุกท่าน ที่ได้ร่วมกันแลกเปลี่ยนความรู้ ให้คำแนะนำ และให้กำลังใจซึ่งกันและกันตลอดระยะเวลาการศึกษา รวมทั้งมีส่วนช่วยให้การดำเนินงานวิจัยฉบับนี้สำเร็จลุล่วงไปได้ด้วยดี

เหนือสิ่งอื่นใด ผู้วิจัยขอกราบขอบพระคุณบิดามารดา ผู้เป็นกำลังใจสำคัญและให้การสนับสนุนผู้วิจัยในทุกๆ ด้าน อย่างต่อเนื่องเสมอมา

สุดท้ายนี้ ผู้วิจัยหวังเป็นอย่างยิ่งว่า วิทยานิพนธ์ฉบับนี้จะเป็นประโยชน์ต่อผู้ที่สนใจ เพื่อนำไปต่อยอดองค์ความรู้และประยุกต์ใช้ให้เกิดคุณค่าในทางวิชาการและการใช้งานต่อไป

ภควัต หาดสมบัติ

TNI

THAI - NICHI INSTITUTE OF TECHNOLOGY

สารบัญ

	หน้า
บทคัดย่อภาษาไทย.....	ง
บทคัดย่อภาษาอังกฤษ.....	จ
กิตติกรรมประกาศ.....	ฉ
สารบัญ.....	ช
สารบัญตาราง.....	ฌ
สารบัญรูป.....	ญ
บทที่	
1 บทนำ.....	1
1.1 ความเป็นมาและความสำคัญของปัญหา.....	1
1.2 วัตถุประสงค์ของงานวิจัย.....	3
1.3 ขอบเขตของการวิจัย.....	3
1.4 ประโยชน์ที่คาดว่าจะได้รับ.....	4
1.5 นิยามศัพท์เฉพาะ.....	4
2 หลักการทฤษฎีและงานวิจัยที่เกี่ยวข้อง.....	5
2.1 ทฤษฎีที่เกี่ยวข้องกับปัญญาประดิษฐ์เชิงสร้างสรรค์ (Generative AI).....	5
2.2 ตัวอย่างเชิงปฏิบัติ (Adversarial Examples).....	13
2.3 แบบจำลองการแพร่กระจาย (Diffusion Model).....	14
2.4 แบบจำลองการแพร่แบบเสถียร (Stable Diffusion).....	17
2.5 ตัวชี้วัดการประเมินผล (Evaluation Metrics).....	18
2.6 งานวิจัยที่เกี่ยวข้อง.....	19
3 ระเบียบและวิธีดำเนินการวิจัย.....	24
3.1 วิเคราะห์ข้อมูลและปัญหาของงานวิจัย.....	24
3.2 การศึกษาข้อมูลและเทคโนโลยีที่ต้องนำมาใช้.....	24
3.3 การออกแบบการทดลอง.....	25

สารบัญ (ต่อ)

บทที่		หน้า
3	3.4 การดำเนินการทดลอง	31
	3.5 การประเมินผลการวิจัย	35
4	ผลการวิจัย.....	36
	4.1 การประเมินคุณภาพของภาพต้นฉบับและภาพที่ผ่านการป้องกัน	36
	4.2 การประเมินผลกระทบของภาพที่ผ่านการป้องกันต่อกระบวนการสร้างภาพ ของโมเดล	41
5	สรุปผลและข้อเสนอแนะ	48
	5.1 สรุปผลการวิจัย.....	48
	5.2 อภิปรายผลการวิจัย.....	49
	5.3 ปัญหาและข้อจำกัด	49
	5.4 ข้อเสนอแนะ	50
	บรรณานุกรม	52
	ภาคผนวก	56
	ภาคผนวก ก. เอกสารประกอบการนำเสนอการสอบป้องกันวิทยานิพนธ์	57
	ประวัติย่อผู้วิจัย.....	104

สารบัญตาราง

ตาราง	หน้า
2.1 สรุปวิธีการหรือเทคนิคที่ใกล้เคียงกับวิธีการของงานวิจัย	21
4.1 ค่าตัวชี้วัด SSIM, PSNR, CLIP Similarity และ LPIPS ของภาพต้นฉบับและภาพที่ ผ่านการป้องกัน.....	36
4.2 ค่าตัวชี้วัด SSIM, PSNR, CLIP Similarity และ LPIPS ของภาพต้นฉบับและภาพที่ ผ่านการป้องกันที่สร้างจากกระบวนการ img2img ของ Stable Diffusion 1.5	42



สารบัญรูป

รูป	หน้า
1.1 รูปที่ถูกสร้างโดย Generative AI.....	1
1.2 รูปอัตราการเติบโตในตลาดโลกของ Generative AI.....	2
2.1 ประเภทของ Generative AI Model ตามสถาปัตยกรรม.....	6
2.2 รูปแบบทั่วไปของการทำงานแบบ Variational Autoencoder (VAE).....	7
2.3 รูปแบบทั่วไปของการทำงานแบบ Normalizing Flow Models.....	8
2.4 รูปแบบทั่วไปของการทำงานแบบ Generative Adversarial Networks.....	9
2.5 รูปแบบทั่วไปของการทำงานแบบ Diffusion Models.....	10
2.6 รูปแบบทั่วไปของการทำงานแบบ Language Models (LMs).....	11
2.7 รูปแบบทั่วไปของการทำงานแบบ Transformers.....	12
2.8 รูปแบบทั่วไปของการทำงานแบบ Hybrid Models.....	13
2.9 รูปแพนด้าที่ถูก Perturbation และใช้ GoogLeNet ทำนายผลลัพธ์ที่ออกมาผิด.....	14
2.10 กระบวนการทำลายและกู้คืนข้อมูลในแบบจำลอง Diffusion Model.....	15
2.11 แนวคิดของแบบจำลอง Diffusion ได้แก่ DDPM, Score-Based SDE และ Conditional Diffusion Model.....	16
3.1 รูปรวมกระบวนการสร้าง Perturbation.....	26
3.2 ขั้นตอนการปรับ Perturbation เชิงละเอียดใน Latent Space.....	30
3.3 ตัวอย่างภาพต้นฉบับ (Original Image) ภาพเป้าหมาย (Target Image) และภาพที่ ผ่านการป้องกัน (Perturbation Image) ในแต่ละหมวดหมู่ ได้แก่ สัตว์ สิ่งของ สถานที่ และบุคคล.....	32
3.4 ตัวอย่างภาพที่สร้างจากภาพต้นฉบับและภาพที่ผ่านการป้องกันภายใต้ค่า Seed ที่แตกต่างกัน.....	33
4.1 การเปรียบเทียบระหว่างภาพต้นฉบับและภาพที่ผ่านการป้องกัน ในกลุ่มตัวอย่าง ประเภทสัตว์.....	38
4.2 การเปรียบเทียบระหว่างภาพต้นฉบับและภาพที่ผ่านการป้องกัน ในกลุ่มตัวอย่าง ประเภทสิ่งของ.....	39
4.3 การเปรียบเทียบระหว่างภาพต้นฉบับและภาพที่ผ่านการป้องกัน ในกลุ่มตัวอย่าง ประเภทสถานที่.....	40

สารบัญรูป (ต่อ)

รูป	หน้า
4.4 การเปรียบเทียบระหว่างภาพต้นฉบับและภาพที่ผ่านการป้องกัน ในกลุ่มตัวอย่าง ประเภทบุคคล	41
4.5 ตัวอย่างผลลัพธ์ของ Stable Diffusion 1.5 ของภาพต้นฉบับ (Original Image) และ ภาพที่ผ่านการป้องกัน (Perturbation Image) ประเภทสัตว์	44
4.6 ตัวอย่างผลลัพธ์ของ Stable Diffusion 1.5 ของภาพต้นฉบับ (Original Image) และ ภาพที่ผ่านการป้องกัน (Perturbation Image) ประเภทสิ่งของ	45
4.7 ตัวอย่างผลลัพธ์ของ Stable Diffusion 1.5 ของภาพต้นฉบับ (Original Image) และ ภาพที่ผ่านการป้องกัน (Perturbation Image) ประเภทสถานที่	46
4.8 ตัวอย่างผลลัพธ์ของ Stable Diffusion 1.5 ของภาพต้นฉบับ (Original Image) และ ภาพที่ผ่านการป้องกัน (Perturbation Image) ประเภทบุคคล	47

TNI

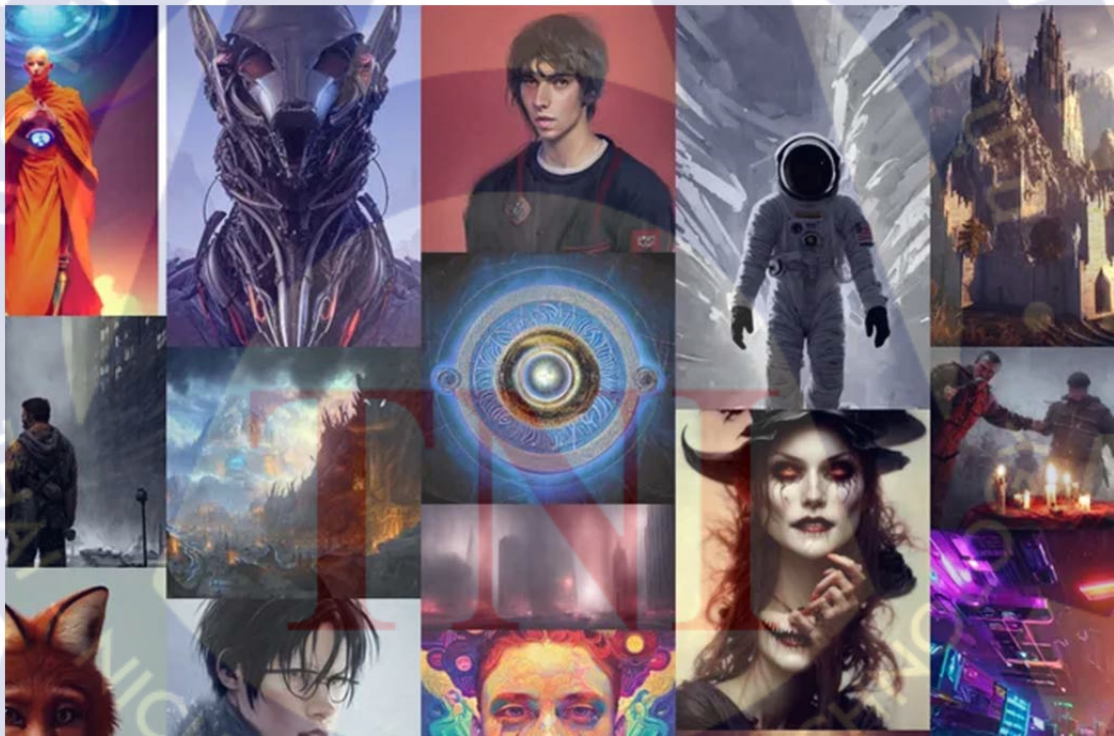
TAI - NICHI INSTITUTE OF TECHNOLOGY

บทที่ 1

บทนำ

1.1 ความเป็นมาและความสำคัญของปัญหา

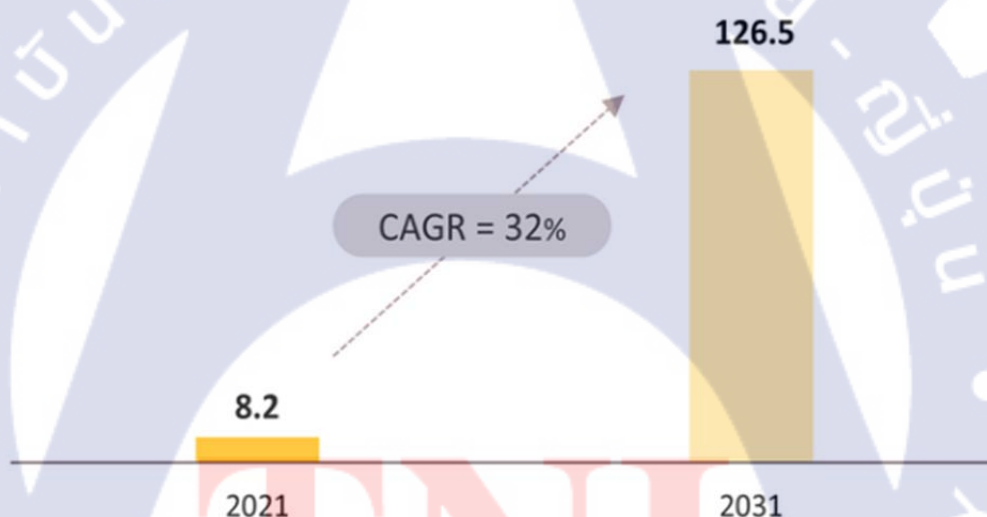
ในปัจจุบัน มีการใช้ปัญญาประดิษฐ์ (AI) รูปแบบใหม่ที่เรียกว่า ปัญญาประดิษฐ์เชิงสร้างสรรค์ (Generative AI) ซึ่งเป็นที่นิยมและมีประโยชน์อย่างมากในการสร้างข้อความ ภาพ วิดีโอ เสียง หรือแม้กระทั่งโค้ดโปรแกรม จากการป้อนข้อความ (Text Prompts) หลักการทำงานของ Generative AI อิงจากการใช้เทคโนโลยีการเรียนรู้ของเครื่อง (Machine Learning) และโครงข่ายประสาทเทียมเชิงลึก (Deep Neural Networks) โดยใช้กระบวนการเรียนรู้ข้อมูลตัวอย่างจำนวนมาก เช่น ข้อความ ภาพ วิดีโอ หรือเสียง จากนั้น AI จะเรียนรู้ รูปแบบ ความสัมพันธ์ โครงสร้างที่ซับซ้อนในข้อมูลเหล่านั้น และใช้ข้อมูลที่เรียนรู้มาเพื่อสร้างเนื้อหาใหม่ที่มีความคล้ายคลึงกับข้อมูลที่เคยถูกฝึกมา [1] ภาพที่ถูกสร้างโดยใช้ Generative AI ซึ่งแสดงให้เห็นในรูปที่ 1.1



รูปที่ 1.1 รูปที่ถูกสร้างโดย Generative AI [1]

ซึ่งกระแสของ Generative AI ได้สร้างความสนใจให้บริษัทยักษ์ใหญ่ทางด้านเทคโนโลยี ไม่ว่าจะเป็น Microsoft ที่เข้าไปลงทุนใน OpenAI รวมมูลค่ากว่า 1.3 หมื่นล้านดอลลาร์สหรัฐเพื่อนำเทคโนโลยีของ OpenAI ซึ่งเป็นผู้พัฒนา ChatGPT มาเพิ่มความสามารถให้กับผลิตภัณฑ์ต่างๆ ของตน นอกจากนี้ Google และ Meta ต่างก็พัฒนา Generative AI ให้พร้อมสำหรับธุรกิจของตนด้วยเช่นกัน ซึ่งไม่เพียงแต่ฝั่งสหรัฐฯ เท่านั้น บริษัทเทคโนโลยียักษ์ใหญ่จากจีนอย่าง Baidu และ Alibaba ต่างก็เร่งพัฒนา Generative AI ของตนเองให้พร้อมแข่งขันกับบริษัทรายใหญ่อื่นๆ ในตลาดโลกด้วยเช่นกัน

Generative AI ทำให้หลายหน่วยงานคาดการณ์ว่ามูลค่าตลาดของเทคโนโลยีนี้จะเติบโตอย่างก้าวกระโดด โดย Allied Market Research เคยประเมินว่าเดิมที Generative AI มีมูลค่าตลาดเพียง 8.2 พันล้านดอลลาร์สหรัฐในปี 2563 (ค.ศ.2021) แต่จะพุ่งเป็น 126.5 พันล้านดอลลาร์สหรัฐในปี 2574 (ค.ศ.2031) หรือคิดเป็นอัตราการเติบโตเฉลี่ยสะสมต่อปี (CAGR) ที่ร้อยละ 32 [2] ภาพอัตราการเติบโตในตลาดโลกของ Generative AI ซึ่งแสดงให้เห็นในรูปที่ 1.2



รูปที่ 1.2 รูปอัตราการเติบโตในตลาดโลกของ Generative AI [2]

แม้ว่า Generative AI จะเป็นเทคโนโลยีที่มีศักยภาพสูงและได้รับการพัฒนาอย่างรวดเร็ว แต่ก็มีข้อกังวลหลายประการ โดยเฉพาะเรื่อง ลิขสิทธิ์ (Copyright) และ จริยธรรม (Ethics) ตัวอย่างเช่น การใช้ข้อมูลที่มีลิขสิทธิ์ในการฝึก AI โดยไม่ได้รับอนุญาต และผลกระทบต่อผู้สร้างเนื้อหาต้นฉบับ เช่น ศิลปินหรือนักเขียน ที่อาจเสียผลประโยชน์จากการที่ AI สามารถสร้างเนื้อหาที่คล้ายกับผลงานของพวกเขาได้โดยง่าย นอกจากนี้ยังมีข้อกังวลเกี่ยวกับความถูกต้องของข้อมูลและการสร้างข้อมูลที่ไม่เป็นความจริง (Misinformation) ที่อาจถูกสร้างขึ้นโดย AI

ในปี 2023 มีการฟ้องร้องหลายคดีที่เกี่ยวข้องกับการใช้ภาพที่มีลิขสิทธิ์เพื่อฝึกโมเดล AI โดยไม่ได้รับอนุญาต หนึ่งในคดีที่เป็นที่รู้จักมากที่สุดคือคดีที่ Getty Images ฟ้องร้อง Stability AI ในข้อหาละเมิดลิขสิทธิ์และใช้ภาพที่มีลายน้ำ (Watermark) ของ Getty โดยที่ไม่ได้นำออกหรือแก้ไขลายน้ำอย่างถูกต้อง โดย Stability AI ถูกกล่าวหาว่าใช้ภาพมากกว่า 12 ล้านภาพ จากคลังของ Getty โดยไม่ได้รับอนุญาต เพื่อฝึกโมเดล Stable Diffusion คดีนี้มีทั้งในสหรัฐอเมริกาและสหราชอาณาจักร และกำลังอยู่ในกระบวนการพิจารณาในศาล [3]

นอกจากนี้ยังมีคดี Andersen v. Stability AI ซึ่งเป็นคดีฟ้องร้องของศิลปินทั้ง 3 คน ได้แก่ Sarah Andersen, Kelly McKernan และ Karla Ortiz ได้ยื่นฟ้อง Generative AI อย่าง Stability AI และ Midjourney รวมถึงแพลตฟอร์มผลงานศิลปะอย่าง DeviantArt ซึ่งสร้างเครื่องมือที่สร้างงานศิลปะด้วย AI ชื่อว่า DreamUp โดยประเด็นสำคัญในการฟ้องร้องคือ การที่เครื่องมือ AI เหล่านี้ถูกฝึกฝนด้วยภาพที่มีลิขสิทธิ์โดยไม่ได้รับอนุญาตจากศิลปินที่เป็นเจ้าของผลงาน ซึ่งส่งผลให้เกิดการละเมิดสิทธิของ "ศิลปินนับล้าน" เนื่องจากภาพเหล่านั้นถูกนำมาใช้ในการสร้างงานศิลปะใหม่ที่เลียนแบบสไตล์ของศิลปินเหล่านั้น โดยศิลปินเหล่านั้นมองว่าการที่ AI สามารถสร้างงานศิลปะใหม่ๆ ได้อย่างไม่จำกัดจะสร้างความเสียหายต่อตลาดงานศิลปะและศิลปิน [4]

จากรายละเอียดที่กล่าวไปข้างต้นผู้วิจัยเล็งเห็นถึงปัญหาการใช้ AI สร้างภาพจากข้อมูลที่มีลิขสิทธิ์และไม่ได้รับการอนุญาตจากศิลปินต้นฉบับ จึงได้นำเสนอวิธีการป้องกันโดยใช้วิธีการแทรกแซงหรือปนเปื้อนข้อมูลในภาพที่ถูกนำไปแก้ไขหรือดัดแปลงโดย Generative AI เพื่อให้สามารถทำลายผลลัพธ์ของโมเดลการสร้างภาพ หรือ ทำให้เกิดความผิดปกติในการสร้างภาพได้

1.2 วัตถุประสงค์ของงานวิจัย

1.2.1 เพื่อพัฒนาอัลกอริทึมการป้องกันภาพที่มีลิขสิทธิ์สำหรับโมเดล Stable Diffusion 1.5 ภายใต้กระบวนการ Image-to-Image (img2img)

1.2.2 เพื่อประเมินความสามารถของวิธีการในการรักษาความใกล้เคียงกับภาพต้นฉบับ ขณะเดียวกันทำให้ผลลัพธ์จากโมเดล Stable Diffusion 1.5 ภายใต้กระบวนการ Image-to-Image (img2img) เกิดความเบี่ยงเบนจากผลลัพธ์ปกติ

1.3 ขอบเขตของการวิจัย

1.3.1 การวิจัยนี้มุ่งเน้นการหาวิธีป้องกันการถูกนำภาพที่มีลิขสิทธิ์ ไปใช้การสร้างภาพใหม่ หรือการดัดแปลงภาพนั้นด้วย Generative AI Model โดยเฉพาะ Stable Diffusion Model 1.5

1.3.2 ประเภทของภาพที่ต้องการป้องกัน คือ สัตว์ สิ่งของ ยานพาหนะ ฯลฯ

1.3.3 ข้อมูลที่ใช้ในการทดลองเป็นชุดข้อมูลสาธารณะ เช่น ImageNet Unsplash และอื่นๆ

1.3.4 การวิจัยนี้จำกัดขอบเขตเฉพาะ Diffusion Model SD 1.5

1.3.5 เมื่อนำภาพที่ถูกป้องกันไปใช้กับ Diffusion Model SD 1.5 จะทำให้เกิดผลลัพธ์ที่มี การเบี่ยงเบนหรือผิดเพี้ยนไปจากผลลัพธ์ปกติ

1.4 ประโยชน์ที่คาดว่าจะได้รับ

สามารถป้องกันการถูกนำภาพที่มีลิขสิทธิ์ไปใช้ใน Generative AI โดยที่ไม่ได้รับการยินยอม จากเจ้าของผลงาน

1.5 นิยามศัพท์เฉพาะ

1.5.1 **Generative AI** หมายถึง ปัญญาประดิษฐ์ที่สามารถสร้างเนื้อหาใหม่ขึ้นมาได้ เช่น ภาพ วิดีโอ ข้อความ หรือดนตรี โดยใช้ข้อมูลที่มีอยู่ในการฝึกเพื่อสร้างสิ่งใหม่ที่ไม่เคยมีมาก่อน

1.5.2 **Model** หมายถึง อัลกอริทึมหรือโครงข่ายประสาทเทียมที่ถูกฝึกให้ทำงาน เช่น การ สร้างภาพหรือข้อความจากข้อมูลที่ถูกฝึก

1.5.3 **Prompt** หมายถึง ข้อความที่ผู้ใช้ป้อนให้ AI เช่น “สร้างภาพของแมว” เพื่อกระตุ้นให้ โมเดลสร้างภาพหรือเนื้อหาตามคำสั่งนั้น

1.5.4 **Diffusion Models** หมายถึง โมเดล Generative ที่สร้างข้อมูลโดยการเพิ่มและลด สัญญาณรบกวน (Noise) ในข้อมูล กระบวนการนี้เริ่มจากการค่อยๆ เพิ่มสัญญาณรบกวนเข้าไปในข้อมูล จนกระทั่งข้อมูลนั้นกลายเป็นรูปแบบที่ต้องการ

บทที่ 2

หลักการทฤษฎีและงานวิจัยที่เกี่ยวข้อง

เพื่อให้บรรลุจุดประสงค์ของงานวิจัยที่ต้องการป้องกันการถูกนำภาพที่มีลิขสิทธิ์ไปใช้ใน Diffusion Model ผู้วิจัยจึงมุ่งเน้นศึกษาทฤษฎีพื้นฐานที่เกี่ยวกับวิธีการต่างๆ ที่จะนำมาใช้ในงานวิจัย โดยแบ่งได้ออกเป็น 5 หัวข้อหลักดังต่อไปนี้

- 2.1 ทฤษฎีที่เกี่ยวข้องกับปัญญาประดิษฐ์เชิงสร้างสรรค์ (Generative AI)
- 2.2 ตัวอย่างเชิงปฏิบัติ (Adversarial Examples)
- 2.3 แบบจำลองการแพร่กระจาย (Diffusion Model)
- 2.4 แบบจำลองการแพร่แบบเสถียร (Stable Diffusion)
- 2.5 ตัวชี้วัดการประเมินผล (Evaluation Metrics)
- 2.6 งานวิจัยที่เกี่ยวข้อง

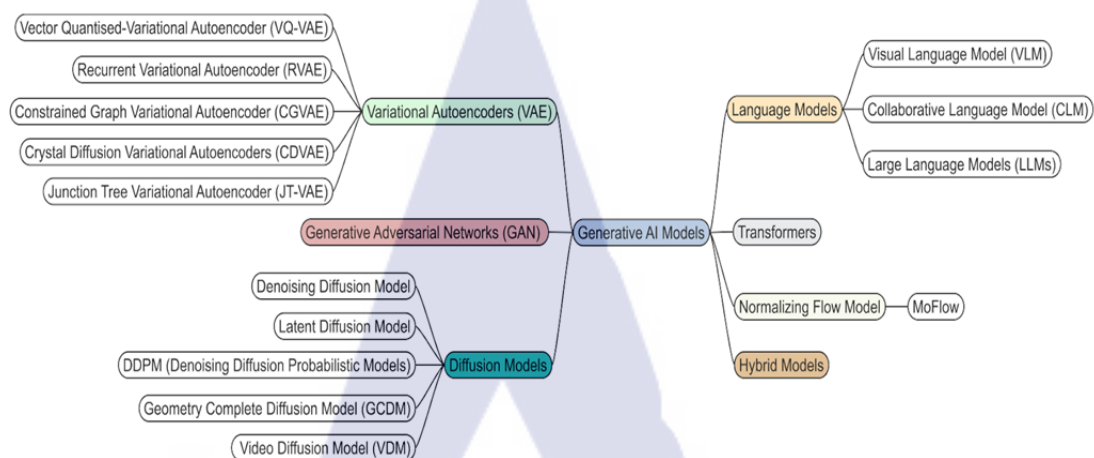
2.1 ทฤษฎีที่เกี่ยวข้องกับปัญญาประดิษฐ์เชิงสร้างสรรค์ (Generative AI)

2.1.1 ความหมายของ Generative AI

Generative AI เป็นระบบปัญญาประดิษฐ์ที่สามารถสร้างข้อความ ภาพ หรือข้อมูลรูปแบบอื่นๆ โดยอาศัยการเรียนรู้ผ่านโมเดลการสร้างข้อมูล (Generative Models) ซึ่งโมเดลเหล่านี้เรียนรู้รูปแบบและโครงสร้างของข้อมูลที่ใช้ฝึก จากนั้นสามารถสร้างข้อมูลใหม่ที่มีลักษณะและคุณสมบัติคล้ายกับข้อมูลต้นแบบที่เคยเรียนรู้มา โดย Generative AI มีความสำคัญมากในปัจจุบันเนื่องจากสามารถสร้างข้อมูลคุณภาพสูงและสมจริง ที่สามารถช่วยในงานต่างๆ เช่น การแปลภาพ การสร้างเนื้อหาจากข้อความ การสร้างภาพจากคำบรรยาย การพัฒนาภาษาสำหรับโมเดลขนาดใหญ่ ไปจนถึงการออกแบบโครงสร้างโปรตีนใหม่ ความสามารถในการเข้าใจและสร้างรูปแบบที่ซับซ้อนจากข้อมูลขนาดใหญ่นี้ทำให้ Generative AI มีบทบาทสำคัญในอุตสาหกรรมหลายด้าน เช่น การแพทย์ การออกแบบ และการสร้างสรรค์ศิลปะ [5]

2.1.2 การจำแนกประเภทของโมเดล Generative AI

โครงสร้างสถาปัตยกรรมของโมเดล Generative AI หมายถึงโครงสร้างพื้นฐานหรือการออกแบบของโมเดล ซึ่งรวมถึงการจัดเรียงเลเยอร์หรือเครือข่ายประสาทเทียมและส่วนประกอบต่างๆ โครงสร้างของโมเดลจะกำหนดวิธีการประมวลผลและการสร้างข้อมูล ซึ่งเป็นปัจจัยสำคัญที่ทำให้โมเดลมีความสามารถเฉพาะสำหรับงานที่จะทำแตกต่างกัน ซึ่งการจำแนกประเภทของโมเดล Generative ตามสถาปัตยกรรมจะแสดง ในรูปที่ 2.1 [5]



รูปที่ 2.1 ประเภทของ Generative AI Model ตามสถาปัตยกรรม [5]

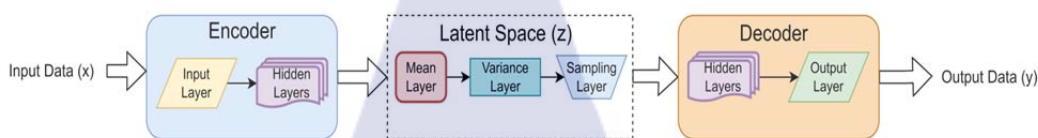
Variational Autoencoder (VAE) เป็นโมเดลการเรียนรู้เชิงลึกแบบ Generative Model ที่ผสมผสานแนวคิดของการอนุมานเชิงความแปรปรวน (Variational Inference) เข้ากับสถาปัตยกรรม Encoder-Decoder เพื่อให้สามารถสร้างข้อมูลใหม่ที่มีลักษณะใกล้เคียงกับข้อมูลจริงที่ใช้ในการฝึก โมเดลส่วน Encoder ทำหน้าที่แปลงข้อมูลอินพุตที่มีมิติสูงให้เป็นการแจกแจงความน่าจะเป็นใน Latent Space ซึ่งมักเป็นการแจกแจงแบบ Gaussian Distribution หลายมิติ ส่วน Decoder จะใช้ข้อมูลที่ถูกรู้สุมตัวอย่างจากการแจกแจงนี้ในการสร้างข้อมูลใหม่ให้มีลักษณะใกล้เคียงกับต้นฉบับ โดยรูปแบบทั่วไปของการทำงานแบบ VAE จะเป็นดังรูปที่ 2.2 การฝึกฝนโมเดล VAE จะอาศัยฟังก์ชันความสูญเสีย (Loss Function) ที่ประกอบด้วยสองส่วนหลัก ได้แก่

- Reconstruction Loss: วัดความแตกต่างระหว่างข้อมูลต้นฉบับกับข้อมูลที่โมเดลสร้างขึ้นใหม่ เพื่อให้ผลลัพธ์ที่สร้างออกมามีความใกล้เคียงกับอินพุตมากที่สุด
- Kullback–Leibler (KL) Divergence: ใช้วัดระยะหรือความเบี่ยงเบนระหว่างการแจกแจงใน latent space กับการแจกแจง Prior ที่ตั้งไว้ล่วงหน้า เพื่อให้การสุ่มตัวอย่างจาก Latent Space มีความเรียบและต่อเนื่อง

VAE มีข้อดีในด้านการสร้างข้อมูลใหม่ที่หลากหลายและมีความคล้ายคลึงกับข้อมูลจริง อย่างไรก็ตาม ข้อมูลที่สร้างขึ้นอาจมีความพร่ามัวและยากต่อการประเมินคุณภาพ นักวิจัยจึงได้พัฒนา VAE รูปแบบต่างๆ เช่น

- VQ-VAE: ใช้ Latent Space แบบไม่ต่อเนื่อง เพื่อเพิ่มความคมชัดของภาพที่สร้างขึ้น
- RVAE: ปรับให้เหมาะสมกับข้อมูลแบบลำดับ (Sequential Data) เช่น ข้อความหรือเสียง
- CGVAE: ออกแบบมาสำหรับข้อมูลในรูปแบบกราฟ (Graph Data)

การพัฒนาเหล่านี้ช่วยให้ VAE สามารถรองรับงานที่ซับซ้อนและสร้างข้อมูลได้อย่างมีประสิทธิภาพมากขึ้น

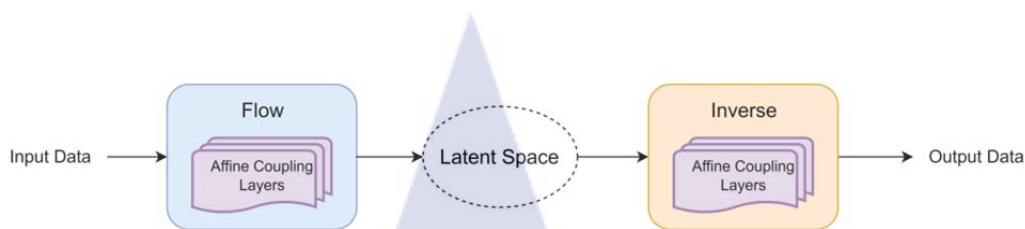


รูปที่ 2.2 รูปแบบทั่วไปของการทำงานแบบ Variational Autoencoder (VAE) [5]

Normalizing Flow Models เป็นโมเดล Generative ที่ใช้การแปลงแบบกำหนดได้และสามารถกลับกันได้ระหว่างข้อมูลดิบและ Latent Space โดยต่างจากโมเดลอื่นอย่าง GANs หรือ VAEs ที่ต้องใช้ Latent Variables โมเดลนี้เน้นการแก้ปัญหการแปลงข้อมูลโดยตรงระหว่างการกระจายข้อมูลสองแบบผ่านการคำนวณ Jacobian Determinant กระบวนการทำงานของ Normalizing Flow คือการนำการแปลงแบบย้อนกลับได้หลายขั้นตอนมาประยุกต์ใช้เพื่อแปลงการกระจายข้อมูลพื้นฐานให้ซับซ้อนขึ้น โดยใช้ Affine Coupling Layers ที่ช่วยให้การแปลงข้อมูลสามารถรักษาคุณสมบัติการกลับกันได้ ทำให้การคำนวณค่าความน่าจะเป็นสามารถทำได้อย่างแม่นยำและรวดเร็ว ซึ่งรูปแบบทั่วไปของการทำงานแบบ Normalizing Flow Models จะแสดงดังรูปที่ 2.3

Affine Coupling Layers มีบทบาทสำคัญในการแปลงข้อมูล โดยจะทำให้ส่วนหนึ่งของข้อมูลยังคงที่ในขณะที่อีกส่วนหนึ่งถูกแปลง การแยกข้อมูลเช่นนี้ทำให้โมเดลสามารถเรียนรู้ความสัมพันธ์ที่ซับซ้อนระหว่างข้อมูลได้ โดยยังคงการคำนวณ Jacobian Determinant อย่างมีประสิทธิภาพ การออกแบบฟังก์ชันที่สามารถย้อนกลับได้และการคำนวณ Jacobian อย่างรวดเร็วเป็นสิ่งสำคัญในการพัฒนาโมเดล Normalizing Flow ให้สามารถประมวลผลข้อมูลที่มีความซับซ้อนได้อย่างแม่นยำ

ข้อได้เปรียบของ Normalizing Flow คือสามารถประเมินความน่าจะเป็นได้อย่างถูกต้องและสุ่มตัวอย่างจากการกระจายข้อมูลที่ซับซ้อนได้อย่างมีประสิทธิภาพ อย่างไรก็ตาม การฝึกฝนโมเดลเหล่านี้อาจใช้ทรัพยากรการคำนวณสูง โดยเฉพาะเมื่อโมเดลมีโครงสร้างที่ลึกหรือข้อมูลมีขนาดใหญ่ เช่น MoFlow ซึ่งเป็นโมเดลที่ใช้การแปลงแบบ Flow เพื่อสร้างแผนภาพโมเลกุลจากการเรียนรู้การแปลงที่สามารถย้อนกลับได้ระหว่างกราฟโมเลกุลและ Latent Representation ของมัน ด้วยการผสมผสานการแปลงแบบต่างๆ ทำให้สามารถสร้างตัวอย่างที่หลากหลายและจับข้อมูลที่ซับซ้อนได้ดีขึ้น

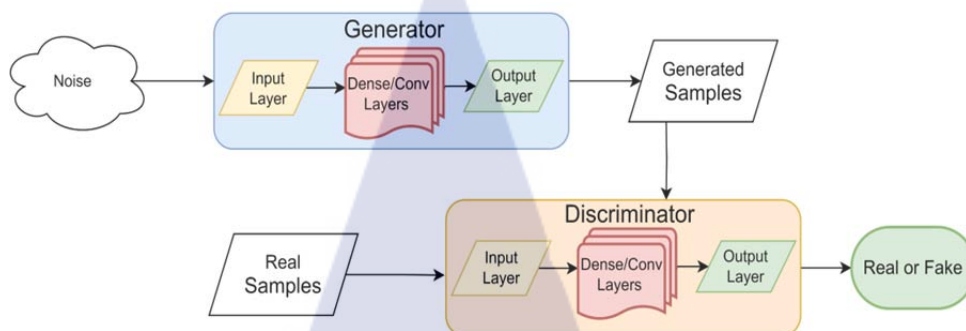


รูปที่ 2.3 รูปแบบทั่วไปของการทำงานแบบ Normalizing Flow Models [5]

Generative Adversarial Networks (GANs) เป็นโมเดล Generative ที่เปิดตัวโดย Ian Goodfellow ในปี 2014 มีโครงสร้างที่ประกอบด้วยสององค์ประกอบหลัก คือ ตัวสร้าง (Generator) และตัวแยกแยะ (Discriminator) โดยทำงานในลักษณะการแข่งขันกัน (Minimax Game) ตัวสร้างจะพยายามสร้างข้อมูลเทียมที่มีลักษณะเหมือนข้อมูลจริง ขณะที่ตัวแยกแยะจะพยายามแยกข้อมูลจริงออกจากข้อมูลเทียม ซึ่งการฝึกฝนนี้จะดำเนินต่อไปจนกว่าตัวแยกแยะจะไม่สามารถแยกแยะข้อมูลจริงจากข้อมูลที่ตัวสร้างสร้างขึ้นได้ ซึ่งรูปแบบทั่วไปของการทำงาน Generative Adversarial Networks จะแสดงดังรูปที่ 2.4

ตัวสร้างจะใช้ตัวแปรจาก Latent Space และสร้างข้อมูลเทียมผ่านเครือข่ายประสาทเทียมที่มีชั้น Dense หรือ Convolutional Layers ในขณะที่ตัวแยกแยะจะรับข้อมูลที่เป็นข้อมูลจริงหรือข้อมูลเทียมที่ตัวสร้างสร้างขึ้น และให้ค่าความน่าจะเป็นว่าข้อมูลนั้นเป็นของจริงหรือเทียม โดยค่าความน่าจะเป็นที่ใกล้ 1 แสดงว่าข้อมูลมาจากข้อมูลจริง และใกล้ 0 แสดงว่าข้อมูลนั้นน่าจะเป็นข้อมูลเทียม

GANs ถูกนำไปใช้ในหลายด้าน เช่น การสร้างภาพที่มีความละเอียดสูง การแปลภาพระหว่างโดเมนต่างๆ การเพิ่มความละเอียดของภาพ และการสร้างภาพจากข้อความ การพัฒนาของ GANs ยังรวมถึงการใช้โครงสร้าง Convolutional เช่นใน DCGAN ที่นำเครือข่าย Convolutional มาใช้ในการสร้างภาพ ทำให้สามารถสร้างภาพที่มีความสมจริงและมีคุณภาพสูงได้ดียิ่งขึ้น GANs ถือเป็นเทคโนโลยีที่สำคัญในการสร้างข้อมูลเชิงสร้างสรรค์และได้เปลี่ยนแปลงหลายวงการในด้านการสร้างภาพและการเรียนรู้ของเครื่อง

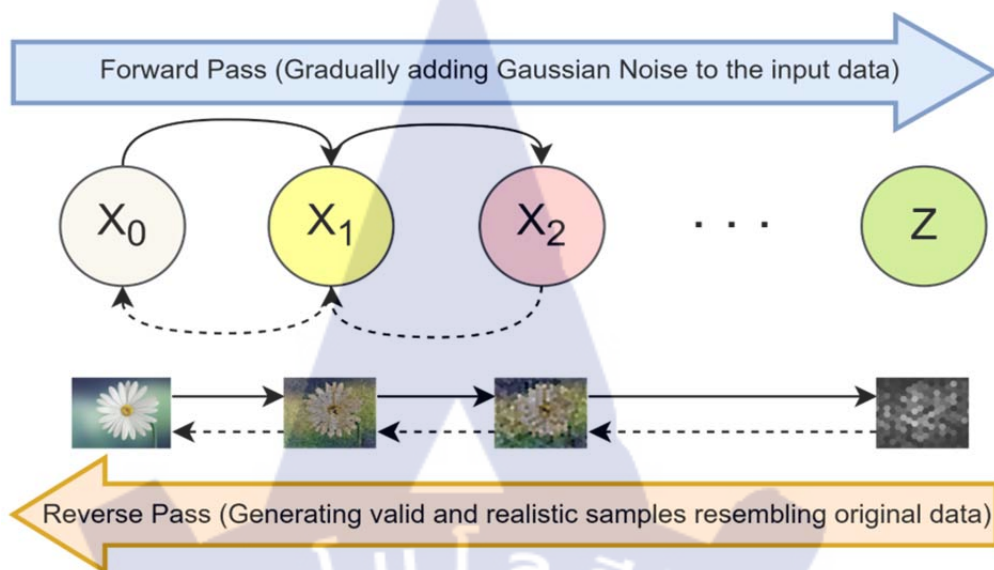


รูปที่ 2.4 รูปแบบทั่วไปของการทำงานแบบ Generative Adversarial Networks [5]

Diffusion Models เป็นโมเดล Generative ที่สร้างข้อมูลโดยการเพิ่มและลดสัญญาณรบกวน (Noise) ในข้อมูล กระบวนการนี้เริ่มจากการค่อยๆ เพิ่มสัญญาณรบกวนเข้าไปในข้อมูลจนกระทั่งข้อมูลนั้นกลายเป็นการกระจายที่ต้องการ เช่น การกระจายแบบ Gaussian หลังจากนั้น โมเดลจะเรียนรู้วิธีการย้อนกลับกระบวนการนี้เพื่อลดสัญญาณรบกวนที่เพิ่มเข้าไป ทำให้สามารถสร้างข้อมูลใหม่ที่มีคุณภาพสูงและคล้ายคลึงกับข้อมูลต้นฉบับได้ซึ่งรูปแบบทั่วไปของการทำงานแบบ Diffusion Models จะแสดงดังรูปที่ 2.5

กระบวนการของ Diffusion Models ประกอบด้วยสองส่วนหลัก คือ การเพิ่มสัญญาณรบกวน (Forward Pass) และการลดสัญญาณรบกวน (Reverse Pass) ในการเพิ่มสัญญาณรบกวน โมเดลจะเพิ่ม Gaussian Noise เข้าไปที่ละชั้น ทำให้ข้อมูลค่อยๆ เสื่อมสภาพ ขณะที่ในการลดสัญญาณรบกวน โมเดลจะเรียนรู้การประมาณความน่าจะเป็นแบบมีเงื่อนไข ซึ่งอธิบายความสัมพันธ์ระหว่างข้อมูลที่ถูกทำลายกับระดับของสัญญาณรบกวน เมื่อโมเดลเรียนรู้กระบวนการนี้แล้ว มันจะสามารถลดสัญญาณรบกวนที่ละชั้นเพื่อสร้างตัวอย่างที่คล้ายคลึงกับข้อมูลต้นฉบับ

Diffusion Models มีรูปแบบย่อยที่แตกต่างกันในการประมวลผล เช่น Denoising Diffusion Probabilistic Models (DDPMs) และ Stochastic Differential Equations (SDEs) แต่ละรูปแบบมีวิธีการที่แตกต่างกันในการจัดการกระบวนการเพิ่มและลดสัญญาณรบกวนเพื่อนำไปสู่การสร้างข้อมูลใหม่ที่มีความสมจริง และมีจุดเด่นในเรื่องการสร้างข้อมูลที่มีรายละเอียดสูงและคมชัด โดยเฉพาะในงานสร้างภาพและสื่อที่ต้องการความสมจริงสูง ทำให้โมเดลนี้ได้รับความนิยมในช่วงหลัง

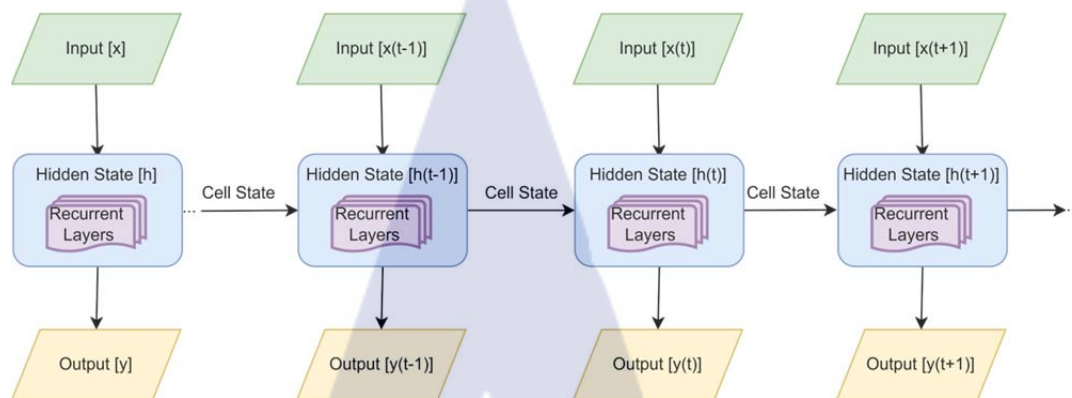


รูปที่ 2.5 รูปแบบทั่วไปของการทำงานแบบ Diffusion Models [5]

Language Models (LMs) คือโมเดลที่ใช้ในการทำความเข้าใจและสร้างข้อความในลักษณะคล้ายมนุษย์ โดยทำงานผ่านหลักการพื้นฐานของการคาดการณ์แบบความน่าจะเป็น โมเดลเรียนรู้รูปแบบและความสัมพันธ์ในลำดับของคำเพื่อประเมินความเป็นไปได้ของคำที่ปรากฏตามบริบทที่กำหนด กระบวนการนี้ช่วยให้โมเดลสามารถสร้างข้อความที่มีความสอดคล้องและตรงตามบริบทได้ ซึ่งรูปแบบทั่วไปของการทำงานแบบ Language Models จะแสดงดังรูปที่ 2.6

Language Models พัฒนาไปอย่างมากในช่วงหลายปีที่ผ่านมา จากการเป็นเครื่องมือสร้างข้อความธรรมดาไปสู่การเป็นเครื่องมือที่สามารถทำความเข้าใจข้อความเชิงลึกได้ โดยใช้การฝึกฝนโมเดลล่วงหน้า (pre-training) เพื่อดึงคุณลักษณะที่ซ่อนอยู่ และใช้คุณลักษณะเหล่านี้ในการทำงานที่ซับซ้อน เช่น การตอบคำถาม การอธิบายเรื่องสามัญสำนึก และการดึงข้อมูลข้อเท็จจริงเกี่ยวกับความสัมพันธ์ระหว่างเอนทิตีต่างๆ

โครงสร้างของ Language Models ประกอบด้วยข้อมูลฝึกฝน โครงสร้างสถาปัตยกรรมของโมเดล และกลไกการอนุมาน (inference) ในการสร้างข้อความ ข้อมูลฝึกฝนเป็นพื้นฐานที่สำคัญในการเรียนรู้รูปแบบและความน่าจะเป็นของภาษา ส่วนโครงสร้างโมเดลอาจประกอบด้วยเครือข่ายประสาทเทียมที่หลากหลาย เช่น Recurrent Neural Networks (RNNs), Long Short-Term Memory (LSTM), Gated Recurrent Units (GRUs) และที่สำคัญคือ Transformers ซึ่งมีความสามารถในการจับความสัมพันธ์ในระยะยาวและเข้าใจบริบทของข้อมูลได้ดีมากขึ้น กลไกการอนุมานจะใช้โมเดลที่ผ่านการฝึกฝนมาเพื่อสร้างข้อความจากการป้อนข้อความเริ่มต้นหรือการทำนายคำที่ขาดหายไป ในบริบทต่างๆ



รูปที่ 2.6 รูปแบบทั่วไปของการทำงานแบบ Language Models (LMs) [5]

Transformers เป็นสถาปัตยกรรมโมเดลที่เปลี่ยนแปลงวงการประมวลผลภาษาธรรมชาติ (NLP) ด้วยการแทนที่ Recurrent Neural Networks (RNNs) แบบเดิมด้วยกลไก Self-Attention ซึ่งช่วยให้โมเดลสามารถโฟกัสไปที่ส่วนต่างๆ ของลำดับข้อมูลได้พร้อมกัน ทำให้การประมวลผลมีประสิทธิภาพมากขึ้น จุดเด่นของ Transformers คือสามารถจับความสัมพันธ์ระยะยาวระหว่างคำหรือตัวอย่างข้อมูลได้โดยการพิจารณาลำดับข้อมูลทั้งหมดในครั้งเดียว ไม่ต้องประมวลผลเป็นลำดับแบบ RNN ซึ่งรูปแบบทั่วไปของการทำงานแบบ Transformers จะแสดงดังรูปที่ 2.7

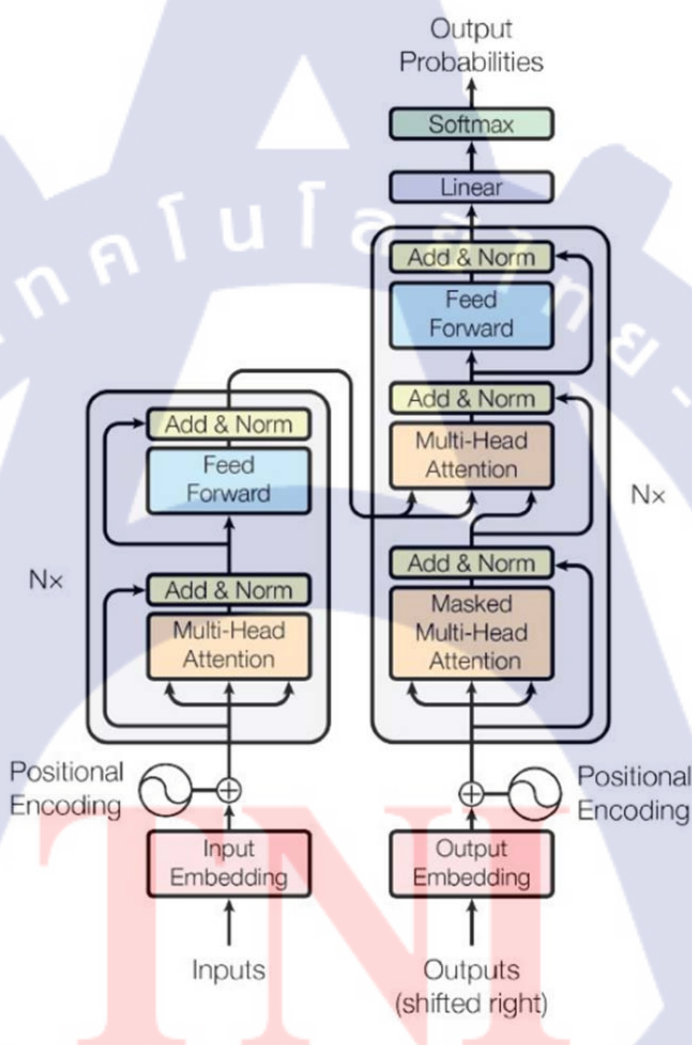
โครงสร้างของ Transformers ประกอบด้วยสองส่วนหลักคือ Encoder และ Decoder โดยทั้งสองส่วนนี้ประกอบด้วยหลายชั้นของ Self-Attention และ Fully Connected Neural Networks ส่วน Encoder ทำหน้าที่ประมวลผลข้อมูลอินพุต ขณะที่ Decoder ทำหน้าที่สร้างข้อมูลเอาต์พุต กลไก Self-Attention ในแต่ละชั้นช่วยให้โมเดลสามารถเลือกสนใจเฉพาะส่วนที่สำคัญในลำดับข้อมูล ซึ่งช่วยให้สามารถจับความสัมพันธ์ที่ซับซ้อนระหว่างคำและทำให้การแปลภาษาและการสร้างข้อความดีขึ้น

การคำนวณ Self-Attention ในโมเดล Transformers ใช้การคำนวณ Dot-Product ระหว่าง Queries (Q), Keys (K), และ Values (V) โดยมีการปรับค่าให้เหมาะสมด้วยตัวประกอบการสเกลและการใช้ฟังก์ชัน Softmax เพื่อให้ได้ค่าความสนใจที่ถูกนำไปคูณกับ Values เพื่อนำไปใช้ในการสร้างผลลัพธ์ในแต่ละขั้นตอน

Transformers ได้รับการพัฒนาไปสู่หลายรูปแบบย่อย เช่น BERT (Bidirectional Encoder Representations from Transformers) ซึ่งใช้ Encoder แบบสองทางในการคาดการณ์คำที่ขาดหายไปจากบริบทและ GPT (Generative Pre-trained Transformer) ที่ใช้ Decoder เพื่อสร้างข้อความตามลำดับโดยใช้การคาดการณ์คำถัดไปจากบริบทก่อนหน้านี้ นอกจากนี้ยังมี T5 (Text-to-Text Transfer

Transformer) ที่ใช้การแปลงข้อความเป็นรูปแบบเดียวกันทั้งอินพุตและเอาต์พุตเพื่อรองรับงานหลากหลายใน NLP

ความก้าวหน้าของ Transformers ช่วยให้สามารถประมวลผลข้อมูลได้อย่างมีประสิทธิภาพ และสามารถทำงานได้หลากหลาย ตั้งแต่การแปลภาษา การสรุปข้อมูล ไปจนถึงการสร้างข้อความ ทำให้กลายเป็นโมเดลหลักที่ได้รับความนิยมในด้านการประมวลผลภาษาธรรมชาติและงานด้าน AI อื่นๆ

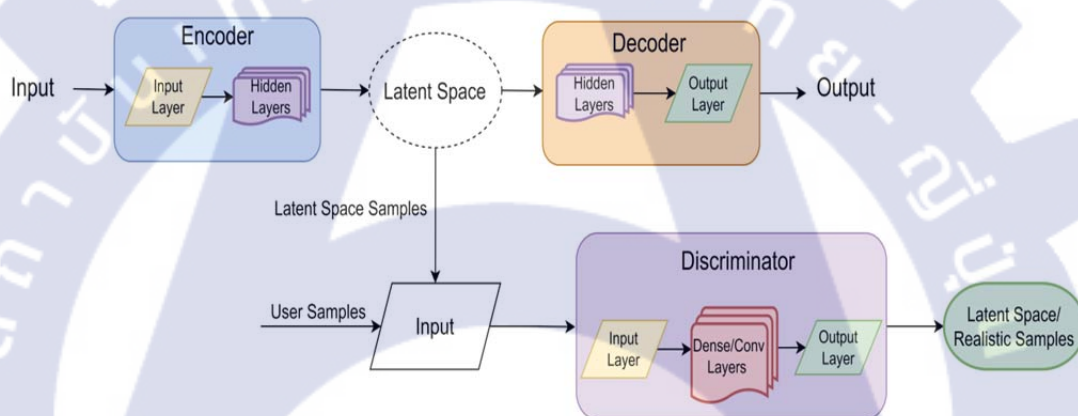


รูปที่ 2.7 รูปแบบทั่วไปของการทำงานแบบ Transformers [5]

Hybrid Models ใน Generative AI เป็นโมเดลที่รวมการทำงานของเทคนิคหรือโครงสร้างหลายแบบเข้าด้วยกันเพื่อเสริมจุดแข็งของแต่ละโมเดลและเพิ่มประสิทธิภาพในการสร้างข้อมูล โดยโมเดลเหล่านี้ถูกพัฒนาเพื่อแก้ไขข้อจำกัดของโมเดลแต่ละประเภทเมื่อใช้งานแยกกัน เช่น การรวม GANs เข้ากับ Denoising Diffusion Models และ Transformers ซึ่งช่วยให้สามารถสร้างตัวอย่างที่มีความ

หลากหลายและคุณภาพสูงขึ้น ด้วยการผสมความสามารถในการสร้างข้อมูลของแต่ละโมเดล ซึ่งรูปแบบทั่วไปของการทำงานแบบ Hybrid Models จะแสดงดังรูปที่ 2.8

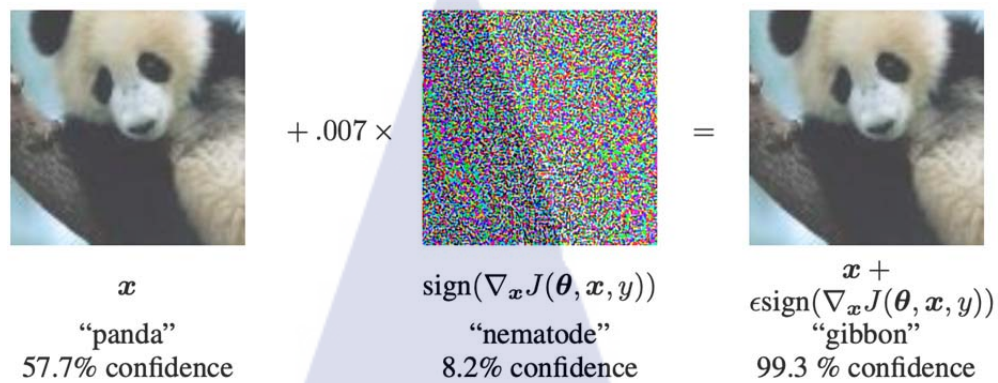
หนึ่งในตัวอย่างของ Hybrid Models คือ Adversarial Autoencoder (AAE) ซึ่งผสมโครงสร้างของ Autoencoders เข้ากับ GANs โดยการใช้ Autoencoder เพื่อเรียนรู้การบีบอัดข้อมูลให้เป็น Latent Representation และใช้ GAN เพื่อทำให้การสร้างตัวอย่างจาก Latent Space มีความสมจริงมากขึ้น นอกจากนี้ยังมีโมเดลเช่น DiffGAN-TTS ที่ผสมผสาน Diffusion Models กับ GANs เพื่อสร้างเสียงที่มีคุณภาพสูง โดยลดจำนวนขั้นตอนในการลดสัญญาณรบกวน ทำให้การสร้างข้อมูลเกิดขึ้นได้รวดเร็วขึ้น โดยรวมแล้ว Hybrid Models มีข้อได้เปรียบในการผสมเทคนิคหลากหลาย ทำให้เกิดความยืดหยุ่นในการออกแบบโมเดลและปรับแต่งให้เหมาะสมกับการใช้งานเฉพาะเจาะจง สามารถสร้างข้อมูลที่มีคุณภาพสูงได้ดีขึ้นเมื่อเทียบกับการใช้โมเดลแต่ละประเภทแยกกัน



รูปที่ 2.8 รูปแบบทั่วไปของการทำงานแบบ Hybrid Models [5]

2.2 ตัวอย่างเชิงปฏิบัติ (Adversarial Examples)

อินพุตที่ถูกปรับแต่งด้วยการเพิ่มสัญญาณรบกวนขนาดเล็ก ซึ่งมักมีความแตกต่างจากข้อมูลต้นฉบับเพียงเล็กน้อยจนแทบไม่สามารถสังเกตเห็นได้ด้วยสายตามนุษย์ แต่การปรับแต่งดังกล่าวสามารถทำให้โมเดลการเรียนรู้เชิงลึก เกิดการทำนายที่คลาดเคลื่อนหรือตัดสินใจผิดพลาดได้อย่างมั่นใจ ลักษณะสำคัญของ Adversarial Example ดังรูปที่ 2.9 ที่แสดงถึงการทำนายผลที่ผิดเพี้ยนของโมเดล



รูปที่ 2.9 รูปแพนด้าที่ถูก Perturbation และใช้ GoogLeNet ทำนายผลลัพธ์ที่ออกมาผิด [6]

ลักษณะสำคัญของ Adversarial Example

- 1) ความแตกต่างจากอินพุตต้นฉบับมักเล็กน้อยจนแทบแยกไม่ออกด้วยสายตามนุษย์ แต่โมเดลกลับให้ผลลัพธ์ที่แตกต่างอย่างสิ้นเชิง
- 2) ตัวอย่างที่สร้างจากโมเดลหนึ่งสามารถนำไปโจมตีโมเดลอื่นที่มีสถาปัตยกรรมหรือข้อมูลฝึกแตกต่างกันได้ ซึ่งสะท้อนถึงการถ่ายโอนของการโจมตี
- 3) โมเดลมักทำนายผลผิดพลาดด้วยระดับความมั่นใจสูง แสดงให้เห็นว่าโมเดลไม่ได้เข้าใจคุณลักษณะที่แท้จริงของข้อมูล แต่เพียงเรียนรู้การประมาณเชิงเส้นของข้อมูลเท่านั้น

Adversarial Example สามารถจำแนกได้หลายลักษณะ เช่น

- ตามความรู้ของผู้โจมตี (Attacker's Knowledge) ได้แก่ White-box ซึ่งผู้โจมตีเข้าถึงโครงสร้างและพารามิเตอร์ของโมเดลได้ทั้งหมด และ Black-box ซึ่งผู้โจมตีไม่รู้รายละเอียดของโมเดล
- ตามวัตถุประสงค์ (Objective) ได้แก่ Targeted Attack ที่ต้องการให้โมเดลทำนายผลผิดเป็นคลาสที่กำหนด และ Untargeted attack ที่เพียงแค่ทำให้โมเดลทำนายผิดจากความจริง
- ตามข้อจำกัดในการรบกวน ได้แก่ การจำกัดขนาดการเปลี่ยนแปลงในเชิงตัวเลข เช่น ใช้ค่าวัดความแตกต่างของพิกเซลที่เล็กมาก
- ตามสภาพแวดล้อม ได้แก่ ตัวอย่างในเชิงดิจิทัลที่สร้างขึ้นในระบบคอมพิวเตอร์ และตัวอย่างเชิงกายภาพที่ถูกนำไปใช้ในโลกรจริง เช่น การติดสติ๊กเกอร์บนป้ายถนนเพื่อหลอกระบบตรวจจับภาพ

2.3 แบบจำลองการแพร่กระจาย (Diffusion Model)

Diffusion Models (DMs) คือแบบจำลองการสร้างข้อมูลที่อาศัยแนวคิด “การทำลายและกู้คืนข้อมูล” กล่าวคือ เริ่มจากนำข้อมูลจริง เช่น ภาพถ่าย ค่อยๆ เติมสัญญาณรบกวน (noise) เข้าไปที่ละน้อย จนข้อมูลกลายเป็น Noise บริสุทธิ์ทั้งหมด นี่เรียกว่า Forward Process จากนั้นโมเดลจะถูกฝึกให้

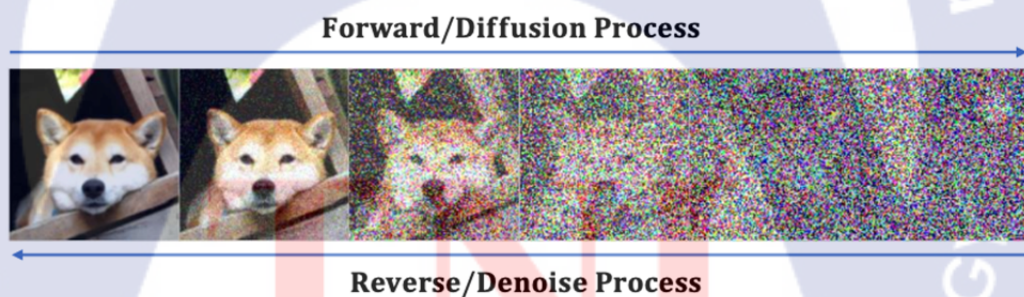
เรียนรู้กระบวนการย้อนกลับ หรือ Reverse Process เพื่อค่อยๆ ลบ Noise ออก จนสามารถสร้างข้อมูลใหม่ที่มีลักษณะใกล้เคียงกับข้อมูลจริงได้ วิธีการนี้ทำให้ Diffusion Models สามารถสร้างภาพที่มีความสมจริงและมีรายละเอียดสูงได้ [7]

1) การทำลายข้อมูล (Forward Process)

ในขั้น forward process โมเดลจะกำหนดวิธีการเติม Noise ให้กับข้อมูลจริง ข้อมูลต้นฉบับ x_0 จะถูกแปลงเป็น $x_1, x_2, \dots, x_T$ ตามลำดับเวลา โดยในแต่ละขั้นจะมีการเพิ่ม Noise เล็กน้อยทำซ้ำไปเรื่อยๆ จนเมื่อถึงขั้นสุดท้าย x_T ข้อมูลจะไม่เหลือโครงสร้างเดิมและมีลักษณะใกล้เคียงกับการแจกแจง Gaussian มาตรฐาน ซึ่งเป็น Distribution ที่ง่ายและเรารู้จักดี จุดประสงค์ของขั้นนี้คือการสร้าง “เส้นทาง” ที่ชัดเจนจากข้อมูลจริงไปยัง Noise ดังที่แสดงในรูปที่ 2.10

2) การกู้คืนข้อมูล (Reverse Process)

เมื่อ Forward Process แปลงข้อมูลจริงให้กลายเป็น Noise ได้แล้ว โมเดล Diffusion จะเรียนรู้การย้อนกระบวนการนี้กลับด้าน โดยเริ่มจาก Noise บริสุทธิ์ที่สุ่มขึ้นมา แล้วค่อยๆ ลบ Noise ออกทีละน้อยในแต่ละขั้นเวลา Reverse Process นี้ใช้เครือข่ายประสาทที่ถูกฝึกให้ทำนายว่า noise ที่เติมเข้าไปคืออะไร แล้วค่อยๆ ลบมันออกไป ผลลัพธ์สุดท้ายคือได้ข้อมูลใหม่ x_0 ที่มีการแจกแจงใกล้เคียงกับข้อมูลจริง วิธีการนี้ไม่ได้ดึงภาพเดิมกลับมา แต่เป็นการสร้างภาพใหม่ที่มีความสมจริงในแบบเดียวกับข้อมูลที่โมเดลเรียนรู้มา ดังที่แสดงในรูปที่ 2.10



รูปที่ 2.10 กระบวนการทำลายและกู้คืนข้อมูลในแบบจำลอง Diffusion Model [8]

3) Denoised Diffusion Probabilistic Model (DDPM)

กระบวนการทำงานเริ่มจากการทำลายข้อมูลจริง x_0 ด้วยการเพิ่ม Noise ทีละขั้นเวลา จนได้ตัวแปร x_T ที่เป็น Gaussian Noise บริสุทธิ์ จากนั้นโมเดลจะเรียนรู้ Reverse Process โดยทำนายสัญญาณรบกวนที่ต้องลบออกในแต่ละขั้น $\epsilon_\theta(x_t, t)$ เพื่อค่อยๆ กู้คืนข้อมูลกลับเป็นตัวอย่างใหม่ที่ใกล้เคียงกับข้อมูลจริง Score-Based Generative Model via Stochastic Differential Equation (SDE) เป็น

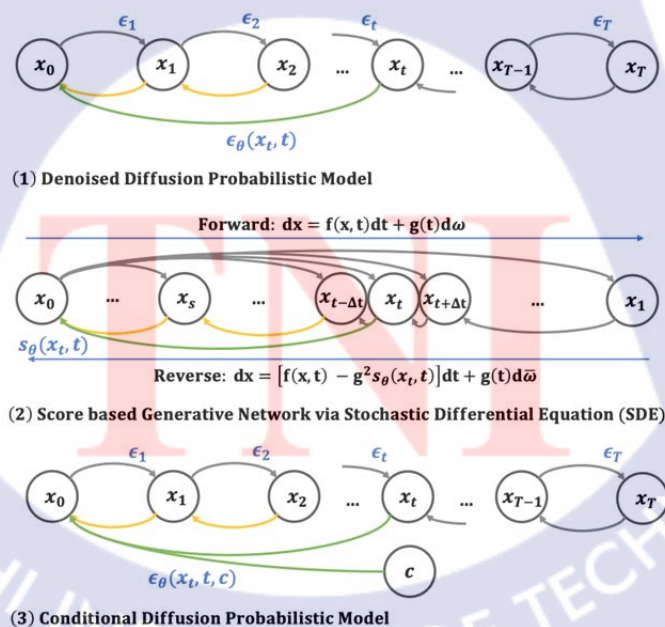
การขยาย Diffusion Models ไปสู่กรอบเวลาเชิงต่อเนื่อง โดยแทน Forward Process ด้วยสมการเชิงอนุพันธ์สุ่ม (SDE) ที่อธิบายการแพร่กระจายของข้อมูลไปเป็น Noise และใช้ Reverse-Time SDE ในการกู้คืนข้อมูลกลับมา โดยอาศัย Score Function $s_\theta(x_t, t)$ ซึ่งประมาณทิศทางของความน่าจะเป็นเพื่อชี้นำทิศทางการสร้างตัวอย่างใหม่

4) Score-Based Generative Network via SDE

เป็นการต่อยอดจากแบบจำลอง DDPM โดยเปลี่ยนกระบวนการเพิ่มและลบสัญญาณรบกวนให้ทำงานอย่างต่อเนื่องในเชิงเวลา โดยใช้สมการเชิงอนุพันธ์สุ่ม (Stochastic Differential Equation; SDE) ในการอธิบายการกระจายของข้อมูลจากภาพจริงไปเป็น Noise และย้อนกลับจาก Noise กลับสู่ภาพใหม่ โมเดลจะเรียนรู้ Score Function เพื่อบอกทิศทางในการลบ Noise ออกจากข้อมูล ทำให้สามารถสร้างภาพได้อย่างราบรื่นและมีความสมจริงมากขึ้น

5) Conditional Diffusion Probabilistic Model

เป็นการต่อยอดแนวคิดของ DDPM โดยเพิ่มตัวแปรเงื่อนไข C เช่น ข้อความหรือป้ายกำกับ เพื่อใช้กำกับ Reverse Process เครื่องมือที่ทำนาย Noise จะพิจารณาทั้งข้อมูลที่มี Noise และตัวแปรเงื่อนไข ทำให้สามารถควบคุมการสร้างข้อมูลใหม่ให้สอดคล้องกับเงื่อนไขที่กำหนด เช่น การสร้างภาพตามคำบรรยายข้อความ ดังที่แสดงในรูปที่ 2.11



รูปที่ 2.11 แนวคิดของแบบจำลอง Diffusion ได้แก่ DDPM, Score-Based SDE และ Conditional Diffusion Model [8]

แบบจำลอง Diffusion ถูกนำไปใช้ในเครื่องมือสร้างสรรค์เชิงพาณิชย์หลายประเภท เช่น DALL-E 2 ของ OpenAI, Imagen ของ Google และ Stable Diffusion ของ Stability AI ซึ่งทั้งหมดอาศัยกระบวนการสร้างภาพจากข้อความ (Text-to-Image) โดยมี Diffusion Model เป็นกลไกหลักในการสร้างภาพที่มีความสมจริงสูง

2.4 แบบจำลองการแพร่แบบเสถียร (Stable Diffusion)

เป็นโมเดลสร้างภาพแบบ Text-to-Image ที่พัฒนามาบนพื้นฐานของ Latent Diffusion Model [9] ซึ่งต่อยอดจากแนวคิด Diffusion Probabilistic Models [10] ที่ได้อธิบายไว้ในหัวข้อก่อนหน้านี้ ความแตกต่างสำคัญคือ Stable Diffusion ดำเนินกระบวนการ Diffusion ใน Latent Space แทนการทำงานใน Pixel Space โดยตรง ทำให้ลดภาระการคำนวณและสามารถสร้างภาพความละเอียดสูงได้อย่างมีประสิทธิภาพ

1) โครงสร้างพื้นฐานของโมเดล

Stable Diffusion ประกอบด้วยองค์ประกอบหลักสามส่วน ได้แก่ Variational Autoencoder (VAE), เครือข่าย UNet สำหรับการลบสัญญาณรบกวน และกลไกการกำกับด้วยข้อความผ่าน Cross-Attention ซึ่งเป็นองค์ประกอบของ Latent Diffusion Model [9]

1.1) Variational Autoencoder (VAE) ทำหน้าที่เข้ารหัสภาพจาก Pixel Space ไปยัง Latent Space ที่มีมิติขนาดเล็กกว่า ซึ่งช่วยลดภาระการคำนวณในกระบวนการ Diffusion จากนั้น Latent Representation ที่ได้จะถูกถอดรหัสกลับเป็นภาพในขั้นตอนสุดท้ายผ่าน Decoder

1.2) UNet Denoising Network เป็นเครือข่ายหลักที่ใช้ในการทำนายสัญญาณรบกวนในแต่ละขั้นเวลา โดยรับ Latent ที่มี Noise เป็น Input และเรียนรู้การลบสัญญาณรบกวนออกตามกระบวนการ Reverse Diffusion ตามแนวคิด Diffusion Probabilistic Model และถูกนำมาประยุกต์ใช้ใน Latent Space ตามแนวทางของ Latent Diffusion Model

1.3) Text Conditioning ผ่าน Cross-Attention ในกระบวนการสร้างภาพจากข้อความ ข้อความจะถูกแปลงเป็น Embedding และนำไปใช้กำกับการทำงานของ UNet ผ่านกลไก Cross-Attention ทำให้โมเดลสามารถสร้างภาพให้สอดคล้องกับคำบรรยายที่กำหนดได้

2) กระบวนการ Sampling ใน Latent Space

กระบวนการสร้างภาพใน Stable Diffusion เริ่มต้นจาก latent noise

$$Z_t \sim N(0,1)$$

ตามแนวคิดของ Diffusion Probabilistic Models จากนั้นโมเดลจะดำเนินกระบวนการลบสัญญาณรบกวนแบบค่อยเป็นค่อยไป (Iterative Denoising) ผ่านเครือข่าย UNet จนได้ Latent Representation ใหม่ ซึ่งจะถูกถอดรหัสกลับผ่าน VAE Decoder เพื่อให้ได้ภาพใน Pixel Space

จำนวนขั้นตอนของการ Sampling (Sampling Steps) มีผลต่อคุณภาพของภาพ โดยจำนวนขั้นตอนที่มากขึ้นมักช่วยให้ภาพมีรายละเอียดที่สมบูรณ์มากขึ้น แต่ใช้เวลาในการคำนวณเพิ่มขึ้นตามลำดับ

ในทางปฏิบัติ Stable Diffusion ยังมีพารามิเตอร์ที่ใช้ควบคุมระดับความสอดคล้องระหว่างภาพที่สร้างกับข้อความ เช่น ค่า Guidance Scale ซึ่งส่งผลต่อความเข้มของการกำกับด้วยข้อความในระหว่างกระบวนการ denoise

3) การกำหนดค่าเริ่มต้นของการสุ่ม (Seed) และความสามารถในการทำซ้ำ (Reproducibility)

การสุ่มค่า Latent Noise ในขั้นเริ่มต้นอาศัยตัวกำเนิดตัวเลขสุ่ม (Random Number Generator) ซึ่งสามารถกำหนดสถานะเริ่มต้นของการสุ่มเพื่อให้เกิดผลลัพธ์ที่ทำซ้ำได้ โดยทั่วไปค่าที่ใช้ควบคุมสถานะเริ่มต้นของการสุ่มนี้เรียกว่า “Seed”

หากกำหนด Seed เดียวกันภายใต้การตั้งค่าพารามิเตอร์อื่นที่เหมือนกัน เช่น จำนวนขั้นตอนการ Sampling หรือค่า Guidance Scale จะทำให้ได้ผลลัพธ์ภาพที่เหมือนกันทุกประการ เนื่องจาก Latent Noise เริ่มต้นมีค่าเดียวกัน ส่งผลให้เส้นทางของกระบวนการ Denoise เป็นไปในทิศทางเดียวกัน ในทางกลับกัน การเปลี่ยนค่า Seed จะทำให้ Latent Noise เริ่มต้นแตกต่างกัน ส่งผลให้เส้นทางของ Reverse Diffusion เปลี่ยนแปลง และนำไปสู่ภาพผลลัพธ์ที่แตกต่างกัน

ดังนั้น Seed แม้จะไม่ใช้พารามิเตอร์ของโมเดลโดยตรง แต่มีบทบาทสำคัญในการควบคุมความสามารถในการทำซ้ำ (Reproducibility) ของกระบวนการสร้างภาพ โดยเป็นตัวกำหนดจุดเริ่มต้นของกระบวนการ Diffusion ใน Latent Space

2.5 ตัวชี้วัดการประเมินผล (Evaluation Metrics)

1) SSIM (Structural Similarity Index Measure)

เป็นตัวชี้วัดคุณภาพของภาพที่ใช้ประเมินความคล้ายคลึงกันระหว่างภาพสองภาพ โดยไม่ได้พิจารณาเฉพาะความแตกต่างเชิงพิกเซลเหมือนตัวชี้วัดแบบดั้งเดิม เช่น MSE หรือ PSNR เท่านั้น แต่ยังคำนึงถึงองค์ประกอบเชิงโครงสร้างของภาพ (Structure) ความสว่าง (Luminance) และความแตกต่างของความเข้มแสง (Contrast) [11]

2) PSNR (Peak Signal-to-Noise Ratio)

เป็นตัวชี้วัดเชิงปริมาณที่ใช้วัดความแตกต่างของค่าพิกเซลระหว่างภาพสองภาพ โดยคำนวณจากค่า Mean Squared Error (MSE) ค่า PSNR ที่สูงหมายความว่าภาพทั้งสองมีความแตกต่างเชิงพิกเซลน้อย ในขณะที่ค่า PSNR ที่ต่ำแสดงว่าความแตกต่างของรายละเอียดภาพเพิ่มขึ้น [12]

3) CLIP

เป็นวิธีการวัดความใกล้เคียงเชิงความหมาย (Semantic Similarity) โดยอาศัยโมเดล CLIP (Contrastive Language-Image Pretraining) ซึ่งถูกฝึกให้แมปข้อมูลภาพและข้อความเข้าสู่ Latent Space ร่วมกัน ภาพและข้อความจะถูกเข้ารหัสเป็นเวกเตอร์เชิงคุณลักษณะ และนำมาคำนวณค่าความใกล้เคียงด้วยวิธีการ เช่น Cosine Similarity หากค่า Similarity มีค่ามาก แสดงว่าโมเดลมองว่าข้อมูลทั้งสองมีความหมายหรือเนื้อหาที่ใกล้เคียงกันในเชิง Semantics [13]

4) LPIPS

LPIPS (Learned Perceptual Image Patch Similarity) เป็นตัวชี้วัด (Metric) ที่ใช้วัดความใกล้เคียงเชิงการรับรู้ (Perceptual Similarity) ระหว่างภาพสองภาพ โดยไม่ได้เปรียบเทียบที่ระดับพิกเซลโดยตรง แต่ใช้ คุณลักษณะที่ได้จากเครือข่ายประสาทเชิงลึก (Deep Neural Network) ที่ผ่านการฝึกมาก่อน (Pre-trained Network) มาประมวลผล แล้วคำนวณระยะห่างระหว่าง Feature Activations ของภาพทั้งสอง หากค่า LPIPS มีค่าต่ำ แสดงว่าภาพทั้งสองมีลักษณะใกล้เคียงกันในมุมมองของการรับรู้ของมนุษย์ [14]

2.6 งานวิจัยที่เกี่ยวข้อง

1) Watermark-embedded Adversarial Examples for Copyright Protection [15]

งานวิจัยนี้เสนอวิธีการป้องกันการลอกเลียนผลงานโดยสร้าง “ตัวอย่างปฏิบัติ” (Adversarial Example) ที่มีการฝังลายน้ำเข้าไปในสัญญาณรบกวน. เมื่อ Diffusion Model พยายามเรียนรู้หรือสังเคราะห์ภาพจากตัวอย่างที่ถูกป้องกันนี้ โมเดลจะถูกบังคับให้สร้างภาพที่มีลายน้ำของผู้สร้างสรรค์ปรากฏขึ้นอย่างชัดเจน ผสมกับสัญญาณรบกวน. วิธีการนี้โดดเด่นในด้านความเร็วในการสร้างตัวอย่างป้องกันที่สูงมากเมื่อเทียบกับวิธีอื่น

2) Nightshade: Prompt-Specific Poisoning Attacks [16]

Nightshade เป็นเทคนิคการ “โจมตีด้วยการวางยาข้อมูล” (Data Poisoning Attack) ที่มีเป้าหมายเพื่อทำให้โมเดล AI ทั้งระบบสร้างภาพตามคอนเซ็ปต์ที่ต้องการผิดพลาดไป โดยอาศัยช่องโหว่ “Concept Sparsity” เพื่อสร้างชุดข้อมูลพิษจำนวนเล็กน้อย (น้อยกว่า 100 ภาพ) ที่เชื่อมโยงคอนเซ็ปต์หนึ่ง (เช่น “สุนัข”) กับรูปภาพของอีกคอนเซ็ปต์หนึ่ง (เช่น “แมว”) เมื่อโมเดลถูกฝึกด้วยข้อมูลพิษ จะ

เกิดการเรียนรู้ที่ผิดพลาด ส่งผลให้เมื่อผู้ใช้ป้อนคำสั่งที่เกี่ยวข้องกับคอนเซ็ปต์เป้าหมาย โมเดลจะสร้างภาพของคอนเซ็ปต์อื่นที่ดูวางยาไว้แทน

3) Anti-DreamBooth: Protecting Users from Personalized Synthesis [17]

งานวิจัยนี้มุ่งเน้นการป้องกันการถูกนำชุดภาพถ่ายส่วนตัวไปใช้สร้างโมเดล AI เฉพาะบุคคล ด้วยเทคนิค DreamBooth. วิธีการคือการเพิ่มสัญญาณรบกวนที่มนุษย์รับรู้ได้ยากเข้าไปในชุดภาพถ่าย เพื่อรบกวนกระบวนการ Fine-Tuning ของเทคนิค DreamBooth โดยเฉพาะผลลัพธ์คือโมเดลไม่สามารถเรียนรู้และจดจำอัตลักษณ์ของบุคคลเป้าหมายได้อย่างสมบูรณ์ ทำให้ภาพบุคคลที่สังเคราะห์ขึ้นมาใหม่มีลักษณะบิดเบี้ยวหรือผิดเพี้ยนจนใช้งานไม่ได้

4) EditShield: Protecting Unauthorized Image Editing [18]

EditShield ถูกออกแบบมาเพื่อป้องกันการแก้ไขภาพจาก AI ที่รับคำสั่งชี้นำเป็นประโยคโดยตรง (Instruction-Guided). เทคนิคนี้จะเพิ่มสัญญาณรบกวนเพื่อรบกวนการเข้ารหัสภาพในปริภูมิแฝง (Latent Representation) ของโมเดล. ซึ่งทำให้ AI ตีความข้อมูลภาพต้นฉบับผิดเพี้ยนไปตั้งแต่แรกเริ่ม เมื่อมีการใช้คำสั่งแก้ไขภาพ ผลลัพธ์ที่ได้จะมีความผิดเพี้ยนสูง เช่น องค์ประกอบในภาพคลาดเคลื่อนหรือวัตถุไม่ตรงกับภาพต้นฉบับ ทำให้การแก้ไขล้มเหลว

5) Adversarial Example Does Good (AdvDM) [19]

งานวิจัยนี้คือการใช้ “ตัวอย่างปฏิปักษ์” เพื่อป้องกันการลอกเลียนแบบสไตล์ภาพวาด โดยสร้างตัวอย่างปฏิปักษ์ที่ทำให้โมเดล AI ไม่สามารถทำความเข้าใจหรือสกัดคุณลักษณะ (Features) ของภาพไปใช้ได้ ผลลัพธ์คือเมื่อ AI พยายามลอกเลียนสไตล์จากภาพที่ถูกป้องกัน จะสร้างภาพออกมาเป็นภาพที่มีสัญญาณรบกวนและความผิดปกติ (Chaotic Content) ซึ่งไม่สามารถนำไปใช้งานได้

6) DiffusionShield: A Watermark for Data Copyright Protection [20]

เทคนิคนี้มีวัตถุประสงค์เพื่อใช้เป็นหลักฐานในการพิสูจน์ว่า AI ถูกฝึกจาก “ชุดข้อมูล” (Dataset) ของเจ้าของลิขสิทธิ์ วิธีการคือฝังลายน้ำที่มองไม่เห็นซึ่งมี “รูปแบบเดียวกัน” (Pattern Uniformity) ในทุกภาพของชุดข้อมูล เพื่อให้ AI เรียนรู้และจดจำลายน้ำนั้นเป็นส่วนหนึ่งของคุณลักษณะข้อมูล ผลลัพธ์คือทุกภาพที่ AI สังเคราะห์ขึ้น จะมีลายน้ำของเจ้าของชุดข้อมูลติดไปด้วย ทำให้สามารถใช้เป็นหลักฐานที่ชัดเจนในการตรวจสอบการละเมิดลิขสิทธิ์ได้

7) Raising the Cost of Malicious AI-Powered Image Editing (Photo Guard) [21]

PhotoGuard เสนอแนวคิด “สร้างภูมิคุ้มกัน” (Immunize) ให้กับภาพถ่าย เพื่อทำให้การแก้ไขดัดแปลงในทางที่ไม่ประสงค์ดีทำได้ยากขึ้น โดยการเพิ่มสัญญาณรบกวนที่มนุษย์รับรู้ได้ยาก ซึ่งจะเข้าไปรบกวนการทำงานของ AI เมื่อมีความพยายามแก้ไขภาพที่ถูกป้องกัน โมเดล AI จะสร้างภาพผลลัพธ์ที่ผิดเพี้ยนและไม่สมจริง ซึ่งเป็นการเพิ่มต้นทุนและอุปสรรคให้กับผู้ไม่หวังดี

รายละเอียดเกี่ยวกับวิธีการ เทคนิค วัตถุประสงค์ และผลลัพธ์ของงานวิจัยที่เกี่ยวข้อง ได้ถูกรวบรวมและสรุปไว้ในตารางที่ 2.1

ตารางที่ 2.1 สรุปวิธีการหรือเทคนิคที่ใกล้เคียงกับวิธีการของงานวิจัย

ชื่อหัวข้อวิจัย	วัตถุประสงค์	วิธีการ/เทคนิค	ผลลัพธ์
Watermark-Embedded Adversarial Examples [15]	ป้องกันการลอกเลียนผลงาน โดยบังคับให้ AI สร้างภาพที่มีลายน้ำของเจ้าของปรากฏขึ้นมา	ใช้วิธีสร้าง ตัวอย่าง ปฏิบัติ (Adversarial Example) ที่มีการฝังข้อมูลลายน้ำ (Watermark) เข้าไปในสัญญาณรบกวน (Perturbation) โดยตรง	เมื่อ Diffusion Model พยายามสังเคราะห์ภาพจากตัวอย่าง ปฏิบัติ โมเดลจะถูกบีบให้สร้าง ภาพที่มีลายน้ำของผู้สร้างสรรค์ปรากฏขึ้นอย่างชัดเจน ซึ่งบ่งชี้ถึงความพยายามในการละเมิดลิขสิทธิ์
Nightshade [16]	“วางยา” (Poisoning) ชุดข้อมูลฝึกสอนของ AI เพื่อให้โมเดลโดยรวมสร้างภาพตามคอนเซ็ปต์เป้าหมายผิดพลาดไป	ใช้วิธีการโจมตีด้วยการวางยาข้อมูล (Data Poisoning Attack) โดยสร้างชุดข้อมูลพิษจำนวนเล็กน้อยที่เชื่อมโยงคอนเซ็ปต์หนึ่ง (เช่น “สุนัข”) กับรูปภาพของอีกคอนเซ็ปต์หนึ่ง (เช่น “แมว”)	โมเดลที่ถูกฝึกด้วยข้อมูลพิษจะ เกิดการเรียนรู้ที่ผิดพลาดเกี่ยวกับคอนเซ็ปต์นั้นๆ ส่งผลให้เมื่อผู้ใช้ป้อนคำสั่ง (Prompt) ที่เกี่ยวข้อง โมเดลจะสร้างภาพของคอนเซ็ปต์อื่นที่ถูกวางยาไว้แทน

ตารางที่ 2.1 สรุปวิธีการหรือเทคนิคที่ใกล้เคียงกับวิธีการของงานวิจัย (ต่อ)

ชื่อหัวข้อวิจัย	วัตถุประสงค์	วิธีการ/เทคนิค	ผลลัพธ์
Anti-DreamBooth [17]	ป้องกันการถูกนำชุดภาพถ่ายส่วนตัว (เช่น ใบหน้า) ไปใช้ สร้างโมเดล AI เฉพาะบุคคล (Personalized Model)	เพิ่มสัญญาณรบกวนที่มนุษย์รับรู้ได้ยากเข้าไปในชุดภาพถ่ายบุคคล เพื่อ รบกวนกระบวนการ Fine-tuning ของเทคนิค Dream Booth โดยเฉพาะ	ทำให้โมเดล ไม่สามารถเรียนรู้และจดจำอัตลักษณ์ของบุคคลเป้าหมายได้อย่างสมบูรณ์ ส่งผลให้ภาพบุคคลที่สังเคราะห์ขึ้นมาใหม่มีลักษณะบิดเบี้ยวจนใช้งานไม่ได้
EditShield [18]	ป้องกันการแก้ไขภาพจาก AI ที่ รับคำสั่งเป็นประโยคโดยตรง (Instruction-guided)	เพิ่มสัญญาณรบกวนเพื่อรบกวนการเข้ารหัสภาพในปริภูมิแฝง (Latent Representation) ทำให้ AI ตีความข้อมูลภาพต้นฉบับผิดเพี้ยนไป	เมื่อมีการใช้คำสั่งแก้ไขภาพ ผลลัพธ์ที่ได้จะมีความผิดเพี้ยนสูง เช่น องค์ประกอบในภาพคลาดเคลื่อน หรือวัตถุไม่ตรงกับภาพต้นฉบับ
Adversarial Example Does Good (AdvDM) [19]	ป้องกันการลอกเลียนแบบสไตล์ โดยทำให้ AI สร้างภาพลอกเลียนแบบไม่สำเร็จ	สร้างตัวอย่างปฏิกิริยาที่ทำให้โมเดล AI ไม่สามารถทำความเข้าใจหรือสกัดคุณลักษณะของภาพไปใช้ได้	โมเดล AI จะสร้างภาพลอกเลียนแบบออกมาเป็นภาพที่มีสัญญาณรบกวนและความผิดปกติ (Chaotic Content) ซึ่งไม่สามารถนำไปใช้งานได้
DiffusionShield [20]	เพื่อใช้เป็นหลักฐานพิสูจน์ว่า AI ถูกฝึกจาก “ชุดข้อมูล” (Dataset) ของเรา	ฝัง ลายน้ำที่มองไม่เห็น ซึ่งมี รูปแบบเดียวกันในทุกภาพของชุดข้อมูล เพื่อให้ AI เรียนรู้และจดจำลายน้ำนั้นเป็นส่วนหนึ่งของคุณลักษณะข้อมูล	ทุกภาพที่ AI สังเคราะห์ขึ้น จะมีลายน้ำของเจ้าของชุดข้อมูลติดไปด้วย ทำให้สามารถใช้เป็นหลักฐานที่ชัดเจนในการตรวจสอบการละเมิดลิขสิทธิ์ได้

ตารางที่ 2.1 สรุปวิธีการหรือเทคนิคที่ใกล้เคียงกับวิธีการของงานวิจัย (ต่อ)

ชื่อหัวข้อวิจัย	วัตถุประสงค์	วิธีการ/เทคนิค	ผลลัพธ์
Raising the Cost of Malicious AI-Powered Image Editing (Photo Guard) [21]	“สร้างภูมิคุ้มกัน” ให้กับภาพถ่าย เพื่อให้การแก้ไขในทางที่ผิดทำได้ยากและมีค่าใช้จ่ายสูงขึ้น	สร้างภูมิคุ้มกันให้ภาพด้วยสัญญาณรบกวนที่มนุษย์รับรู้ได้ยาก ซึ่งจะเข้าไป รบกวนกระบวนการทำงานภายในของ AI	เมื่อมีการแก้ไขภาพโมเดล AI จะสร้างภาพผลลัพธ์ที่ผิดเพี้ยนและมีคุณภาพต่ำ

จากการศึกษางานวิจัยที่เกี่ยวข้องพบว่า งานส่วนใหญ่มุ่งเน้นการฝังลายน้ำ การวางยาข้อมูล หรือการป้องกันการแก้ไขภาพตามคำสั่ง แม้ว่าบางงานจะใช้เทคนิคการรบกวนในปริภูมิแฝง (Latent Space) และทดสอบบนโมเดล Stable Diffusion เช่นเดียวกัน แต่ยังไม่ได้มุ่งเน้นการควบคุมพฤติกรรมของโมเดลในบริบทของกระบวนการ Image-to-Image โดยเฉพาะ ดังนั้น งานวิจัยฉบับนี้จึงเสนอแนวทางการปรับค่า Perturbation ใน Latent Space ภายใต้กรอบการเพิ่มประสิทธิภาพเชิง Optimization โดยออกแบบฟังก์ชันสูญเสีย (Loss Function) หลายองค์ประกอบ เพื่อสร้างสมดุลระหว่างการรักษาความใกล้เคียงกับภาพต้นฉบับ และการทำให้ผลลัพธ์จากกระบวนการ img2img แตกต่างจากกรณีที่ใช้ภาพต้นฉบับโดยตรง

บทที่ 3

ระเบียบและวิธีดำเนินการวิจัย

จากการศึกษางานวิจัย แนวคิด และทฤษฎีที่เกี่ยวข้อง ผู้วิจัยได้นำองค์ความรู้มาประยุกต์ใช้เพื่อพัฒนากระบวนการและวิธีการที่สามารถป้องกันไม่ให้ภาพที่มีลิขสิทธิ์ถูกนำไปใช้งานหรือถูกดัดแปลงภายในกระบวนการของ Diffusion Model โดยมุ่งเน้นไปที่การวิเคราะห์วิธีการป้องกันที่มี และออกแบบแนวทางใหม่ที่สามารถป้องกันการถูกนำภาพไปใช้โดยไม่ได้รับอนุญาตได้อย่างมีประสิทธิภาพ โดยมีขั้นตอนการดำเนินงานดังนี้

- ขั้นตอนที่ 1 วิเคราะห์ข้อมูลและปัญหาของงานวิจัย
- ขั้นตอนที่ 2 การศึกษาข้อมูลและเทคโนโลยีที่ต้องนำมาใช้
- ขั้นตอนที่ 3 การออกแบบการทดลอง
- ขั้นตอนที่ 4 การดำเนินการทดลอง
- ขั้นตอนที่ 5 การประเมินผลการวิจัย

3.1 วิเคราะห์ข้อมูลและปัญหาของงานวิจัย

ผู้วิจัยได้ทำการศึกษาปัญหาที่เกิดขึ้นในปัจจุบัน ซึ่งเป็นปัญหาเกี่ยวกับการนำปัญญาประดิษฐ์เชิงสร้างสรรค์ (Generative Model) โดยเฉพาะ Diffusion Model มาใช้สร้างและดัดแปลงภาพอย่างแพร่หลาย ปัญหาสำคัญที่พบ คือ การนำภาพที่มีลิขสิทธิ์มาใช้ โดยไม่ได้รับอนุญาต ส่งผลให้เกิดประเด็นด้านการละเมิดลิขสิทธิ์ และสร้างความกังวลต่อเจ้าของผลงานที่เป็นผู้สร้างสรรค์ต้นฉบับ

จากการศึกษาปัญหาที่เกิดขึ้นรวมถึงแนวทางในการแก้ไขปัญหา ทำให้ผู้วิจัยเลือกทำงานวิจัยในหัวข้อ “การออกแบบและพัฒนาวิธีการป้องกันภาพที่มีลิขสิทธิ์ถูกใช้ใน Generative AI”

3.2 การศึกษาข้อมูลและเทคโนโลยีที่ต้องนำมาใช้

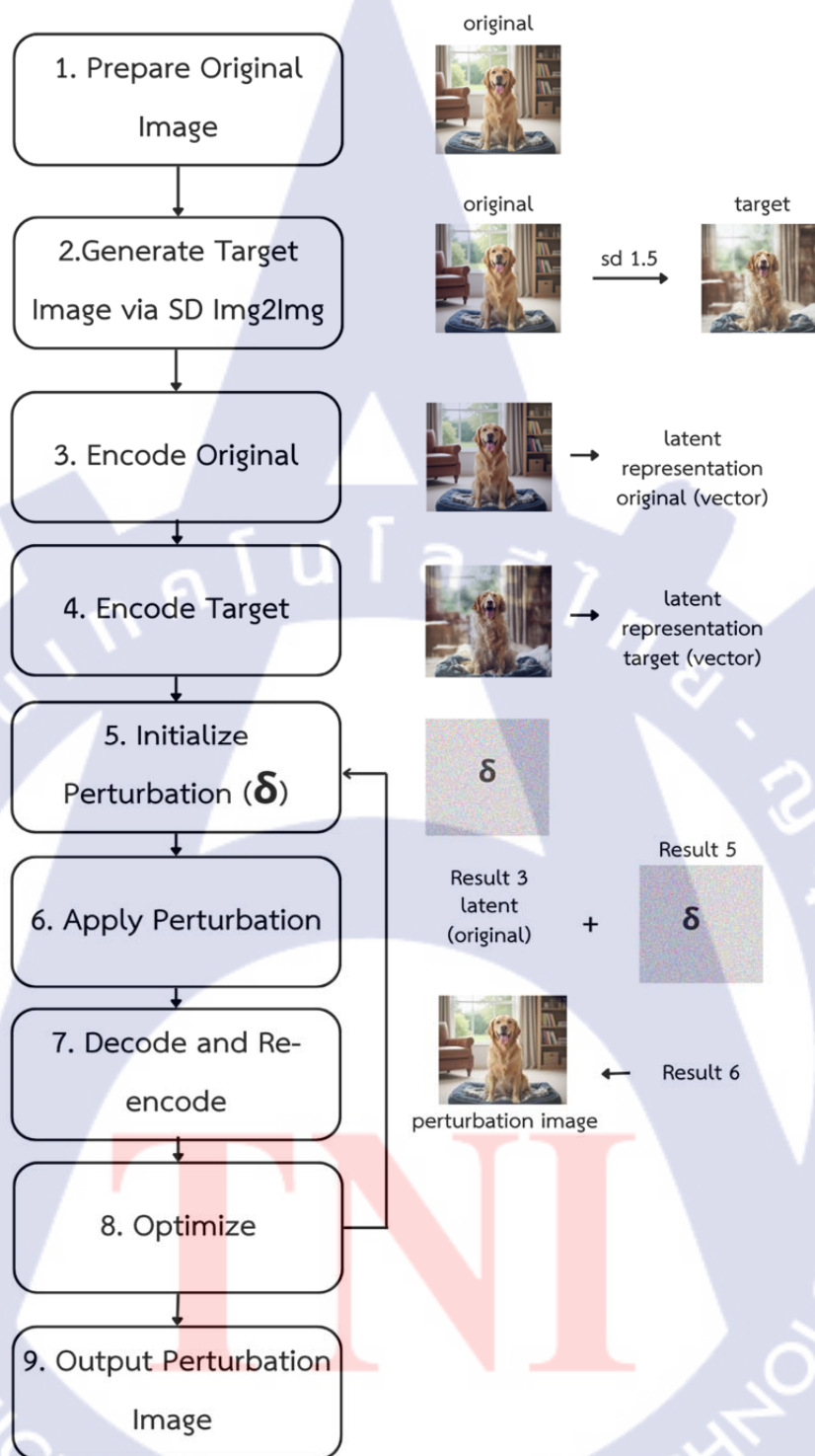
ผู้วิจัยได้ทำการศึกษาข้อมูลและเทคโนโลยีที่เกี่ยวข้องกับการสร้างและการป้องกันภาพในบริบทของปัญญาประดิษฐ์เชิงสร้างสรรค์ (Generative AI) โดยเฉพาะ Diffusion Model ซึ่งเป็นโมเดลที่ได้รับความนิยมอย่างมากในการสร้างและดัดแปลงภาพ เนื่องจากสามารถสร้างภาพที่มีความสมจริงและมีคุณภาพสูง ปัญหาที่เกิดขึ้นตามมา คือ การที่ภาพที่มีลิขสิทธิ์ถูกนำไปใช้งานหรือนำไปดัดแปลงโดยไม่ได้รับอนุญาต ซึ่งก่อให้เกิดความเสี่ยงด้านการละเมิดสิทธิ์ทางปัญญา เทคโนโลยีที่ผู้วิจัยศึกษาเพื่อนำมาใช้ในการแก้ปัญหานี้แบ่งออกได้เป็น 3 ส่วนหลัก ได้แก่ 1) เทคโนโลยี Diffusion Model ศึกษาโครงสร้างและกระบวนการทำงานของ Diffusion Model ทั้งในส่วน Forward Process (การเพิ่มสัญญาณรบกวน

เข้าไปในข้อมูล) และ Reverse Process (การกำจัดสัญญาณรบกวนเพื่อสร้างภาพใหม่) เพื่อทำความเข้าใจจุดที่สามารถแทรกแซงหรือป้องกันไม่ให้ภาพที่มีลิขสิทธิ์ถูกนำไปใช้งานโดยไม่ได้รับอนุญาต 2) เทคนิคการสร้าง Perturbation (การรบกวนเชิงคำนวณ) ที่สามารถซ่อนอยู่ในภาพโดยไม่กระทบต่อการมองเห็นของมนุษย์ แต่สามารถป้องกันไม่ให้ภาพถูกนำไปใช้ในโมเดลการสร้างภาพใหม่ๆ ได้ 3) Evaluation Metrics ใช้ตัวชี้วัดและเกณฑ์การประเมิน ได้แก่ SSIM, PSNR, CLIP, LPIPS เพื่อประเมินประสิทธิภาพของเทคนิคการป้องกันและตรวจสอบผลกระทบที่เกิดขึ้นกับคุณภาพของภาพ

3.3.การออกแบบการทดลอง

จากการศึกษาข้อมูลและเทคโนโลยีที่เกี่ยวข้อง ผู้วิจัยได้ออกแบบระบบการทดลองเพื่อทดสอบประสิทธิภาพของวิธีการป้องกันภาพที่มีลิขสิทธิ์ไม่ให้ถูกนำไปดัดแปลงด้วย Stable Diffusion Model 1.5

ภาพรวมของกระบวนการสร้างการรบกวน (Perturbation) เพื่อให้โมเดลตีความภาพเปลี่ยนไปจากเดิม โดยจะทำการเข้ารหัสภาพต้นฉบับและภาพเป้าหมายเข้าสู่ Latent Space เพื่อนำมาเปรียบเทียบและหาค่าการรบกวนที่เหมาะสม (δ) จากนั้นจึงปรับค่า δ ซ้ำหลายรอบเพื่อให้เมื่อรวมกลับเข้ากับภาพต้นฉบับแล้ว โมเดลสร้างภาพที่มีความหมายหรือรายละเอียดเบี่ยงเบนจากต้นฉบับได้ตามที่ต้องการ ซึ่งขั้นตอนการออกแบบ Perturbation จะเป็นดังรูปที่ 3.1



รูปที่ 3.1 รูปรวมกระบวนการสร้าง Perturbation

1. Prepare Data

ในขั้นตอนแรก จะต้องทำการเตรียมข้อมูลที่ใช้สำหรับการทดลอง โดยข้อมูลดังกล่าวก็คือ ภาพต้นฉบับ (Original) ที่เราอยากจะป้องกัน

2. Generate Target Image via SD Img2Img

นำภาพต้นฉบับเข้าสู่โมเดล Stable Diffusion 1.5 ในโหมด Image-to-Image พร้อมกำหนด Prompt และพารามิเตอร์ต่างๆ เพื่อสร้างภาพเป้าหมาย (Target Image) ซึ่งเป็นภาพที่มีลักษณะ ผิดเพี้ยนจากต้นฉบับ ภาพนี้จะถูกใช้เป็นแนวทางในการกำหนดทิศทางของการปรับ Latent Representation ในกระบวนการสร้าง Perturbation

3. Encode Original

นำภาพต้นฉบับเข้าสู่กระบวนการเข้ารหัส (Encoding) ผ่านตัวเข้ารหัสแบบ VAE (Encoder) เพื่อแปลงภาพให้อยู่ในรูปแบบของข้อมูลที่ถูกบีบอัด โดยเฉพาะ “คุณลักษณะสำคัญ” (Features) ของภาพ เรียกว่า Latent Representation ซึ่งเป็นการแทนค่าภาพในเชิงคุณลักษณะ ที่สามารถนำไปใช้ ประมวลผลต่อได้อย่างมีประสิทธิภาพ

4. Encode Target

ภาพเป้าหมาย (Target) จะถูกเข้ารหัสด้วยวิธีการเดียวกันเพื่อให้ได้ Latent Representation ที่สอดคล้องกับข้อมูลเป้าหมาย การแทนค่าดังกล่าวสามารถใช้เปรียบเทียบกับ Latent ของภาพต้นฉบับ เพื่อเป็นแนวทางในการสร้าง perturbation ที่เหมาะสม

5. Initialize Perturbation (δ)

ในขั้นตอนนี้จะทำการกำหนดตัวแปรสำหรับสร้างค่าการรบกวนเชิงแฝง โดยกำหนดตัวแปร u ซึ่งมีขนาดเท่ากับ Latent ของภาพต้นฉบับ และถูกใช้เป็นพารามิเตอร์สำหรับสร้าง Perturbation (δ) ใน Latent Space พร้อมทั้งกำหนดขอบเขตความแรงสูงสุดของ Perturbation เพื่อป้องกันไม่ให้เกิดการ ปรับแต่งส่งผลกระทบต่อภาพในเชิงการมองเห็นมากเกินไป นอกจากนี้ยังมีการกำหนด Adam Optimizer และอัตราการเรียนรู้ เพื่อใช้ปรับปรุงค่า u แบบวนซ้ำ ก่อนนำมาคำนวณเป็น Perturbation (δ) และ รวมเข้ากับ Latent ของภาพต้นฉบับในขั้นตอนถัดไป

6. Apply Perturbation

เมื่อได้ Perturbation (δ) แล้ว ระบบจะนำค่าดังกล่าวไปรวมกับ Latent ของภาพต้นฉบับ ใน Latent Space เพื่อสร้าง Latent ใหม่ที่ถูกปรับแต่ง ซึ่ง Latent ใหม่นี้เป็นตัวแทนของภาพที่มีการ ผัง Perturbation แล้ว

7. Decode and Re-encode

จากนั้น Latent ที่ถูกปรับแต่งจะถูกถอดรหัส (Decoding) ผ่าน VAE Decoder เพื่อแปลงกลับมาเป็นภาพในระดับพิกเซล ได้เป็นภาพ Perturbation Image ซึ่งภาพดังกล่าวจะถูกนำกลับเข้า VAE Encoder อีกครั้ง เพื่อแปลงเป็น Latent Representation สำหรับใช้ในการคำนวณ Loss และปรับค่า Perturbation ในรอบถัดไป

8. การคำนวณ Losses

ระบบจะทำการประเมินคุณภาพของภาพที่ถูกปรับแต่งในแต่ละรอบ และปรับปรุงค่าการรบกวนเชิงแฝง (Perturbation, δ) อย่างต่อเนื่องผ่านกระบวนการวนซ้ำ (Iterative Optimization) เพื่อให้ได้ Perturbation ที่เหมาะสม โดยนำค่า Loss ทุกตัวมารวมเป็น Total Loss จากนั้นคำนวณ Gradient ด้วย Backpropagation และใช้ Adam Optimizer ปรับค่าตัวแปร u เพื่ออัปเดต Perturbation ในรอบถัดไป

- Adversarial Loss บังคับให้ latent ของภาพที่ถูกปรับแต่งใกล้เคียงกับ Latent ของภาพเป้าหมาย
- Latent-to-Target Loss ทำหน้าที่วัดระยะห่างระหว่าง Latent ใหม่ที่ถูกปรับกับ Latent ของภาพเป้าหมาย
- Keep loss ใช้ควบคุมไม่ให้ Latent ใหม่เบี่ยงเบนออกจาก Latent ของภาพต้นฉบับมากเกินไป
- Pixel Loss ใช้เพื่อวัดความแตกต่างระหว่างภาพใหม่และภาพต้นฉบับในระดับพิกเซล
- Perturbation Loss ซึ่งควบคุมความแรงของ Perturbation โดยอนุญาตให้ภาพเปลี่ยนแปลงได้ภายในขอบเขตที่กำหนด

เพื่อให้กระบวนการปรับค่า Perturbation เป็นไปอย่างมีระบบ ฟังก์ชันความสูญเสียรวม (Total Loss) ถูกกำหนดให้อยู่ในรูปแบบการถ่วงน้ำหนักขององค์ประกอบต่างๆ ดังสมการต่อไปนี้

$$L_{\text{total}} = \lambda_{\text{adv}} L_{\text{adv}} + \lambda_{\text{lat}} L_{\text{lat}} + \lambda_{\text{keep}} L_{\text{keep}} + \lambda_{\text{pix}} L_{\text{pix}} + \lambda_{\text{pert}} L_{\text{pert}}$$

โดยที่ λ เป็นค่าน้ำหนัก (Weight) สำหรับควบคุมความสำคัญของ Loss แต่ละองค์ประกอบ ซึ่งมีบทบาทในการสร้างสมดุลระหว่างการรบกวนเชิงแฝงและการรักษาคุณภาพของภาพต้นฉบับ ในกรทดลองนี้ กำหนดค่าน้ำหนักของแต่ละองค์ประกอบดังนี้

$$\lambda_{adv} = 1.2$$

$$\lambda_{lat} = 0.5$$

$$\lambda_{keep} = 0.5$$

$$\lambda_{pix} = 0.3$$

$$\lambda_{pert} = 10.0$$

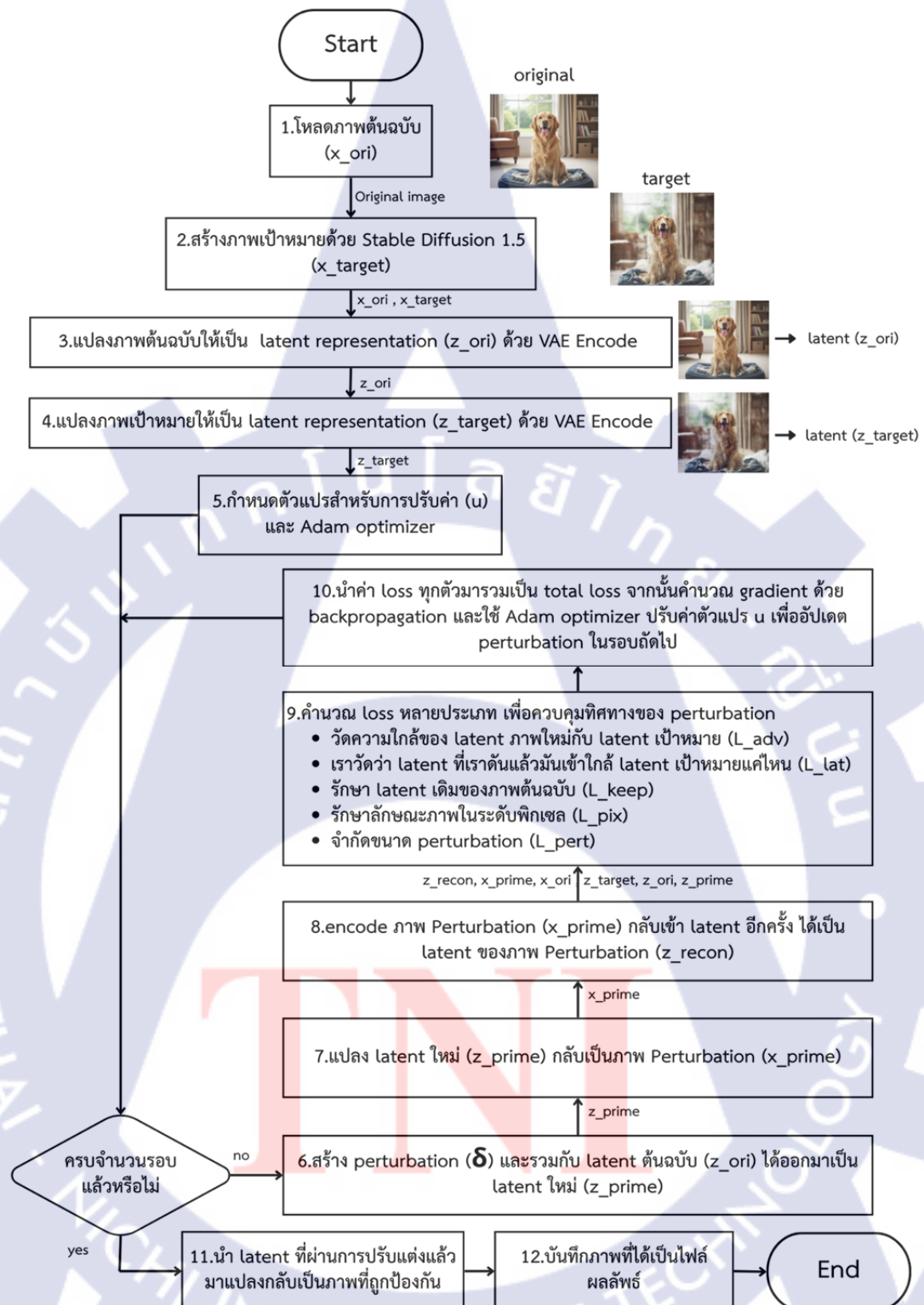
การกำหนดค่าน้ำหนักดังกล่าวมีวัตถุประสงค์เพื่อให้เกิดความสมดุลระหว่างการผลักดัน Latent Representation ไปยังทิศทางของภาพเป้าหมาย (ผ่าน L_{adv} และ L_{lat}) การควบคุมไม่ให้เกิดการเบี่ยงเบนจากภาพต้นฉบับมากเกินไป (ผ่าน L_{keep} และ L_{pix}) และการจำกัดขนาดของ Perturbation ให้อยู่ภายในขอบเขตที่กำหนด (ผ่าน L_{pert}) ซึ่งเป็นกลไกสำคัญที่ช่วยป้องกันไม่ให้อาพเกิดความผิดเพี้ยนในเชิงการมองเห็นมากเกินไป

ทั้งนี้ ค่าน้ำหนักดังกล่าวไม่จำเป็นต้องมีผลรวมเท่ากับ 1 เนื่องจากเป็นเพียงตัวกำหนดความสำคัญของ Loss แต่ละตัว โดยสิ่งที่สำคัญคือสัดส่วนระหว่างค่าน้ำหนักแต่ละองค์ประกอบ มากกว่าผลรวมทั้งหมด ทั้งนี้ ค่าที่ใช้ได้มาจากการทดลองปรับหลายครั้งเพื่อให้ได้ความสมดุลที่เหมาะสม

9. Output Perturbation Image

หลังจากกระบวนการปรับปรุง Perturbation เสร็จสิ้นแล้ว ระบบจะนำ Perturbation สุดท้ายที่ได้ไปประยุกต์ใช้กับ Latent ของภาพต้นฉบับ และถอดรหัสกลับมาเป็นภาพในระดับพิกเซลอีกครั้ง ผลลัพธ์ที่ได้คือ ภาพที่ผ่านการป้องกัน (Perturbation Image) ซึ่งมีลักษณะใกล้เคียงกับภาพต้นฉบับในเชิงการมองเห็น แต่มีการฝัง Perturbation ที่สามารถส่งผลกระทบต่อการทำงานของโมเดล Diffusion ในขั้นตอนการสร้างภาพต่อไป

จากรูปที่ 3.1 ซึ่งแสดงภาพรวมของกระบวนการสร้าง Perturbation และเพื่ออธิบายรายละเอียดของขั้นตอนการปรับ Perturbation ใน Latent Space อย่างชัดเจนยิ่งขึ้น จึงนำเสนอขั้นตอนเชิงละเอียดดังแสดงในรูปที่ 3.2 ขั้นตอนการปรับ Perturbation เชิงละเอียดใน Latent Space



รูปที่ 3.2 ขั้นตอนการปรับ Perturbation เจริญละเอียดใน Latent Space

3.4 การดำเนินการทดลอง

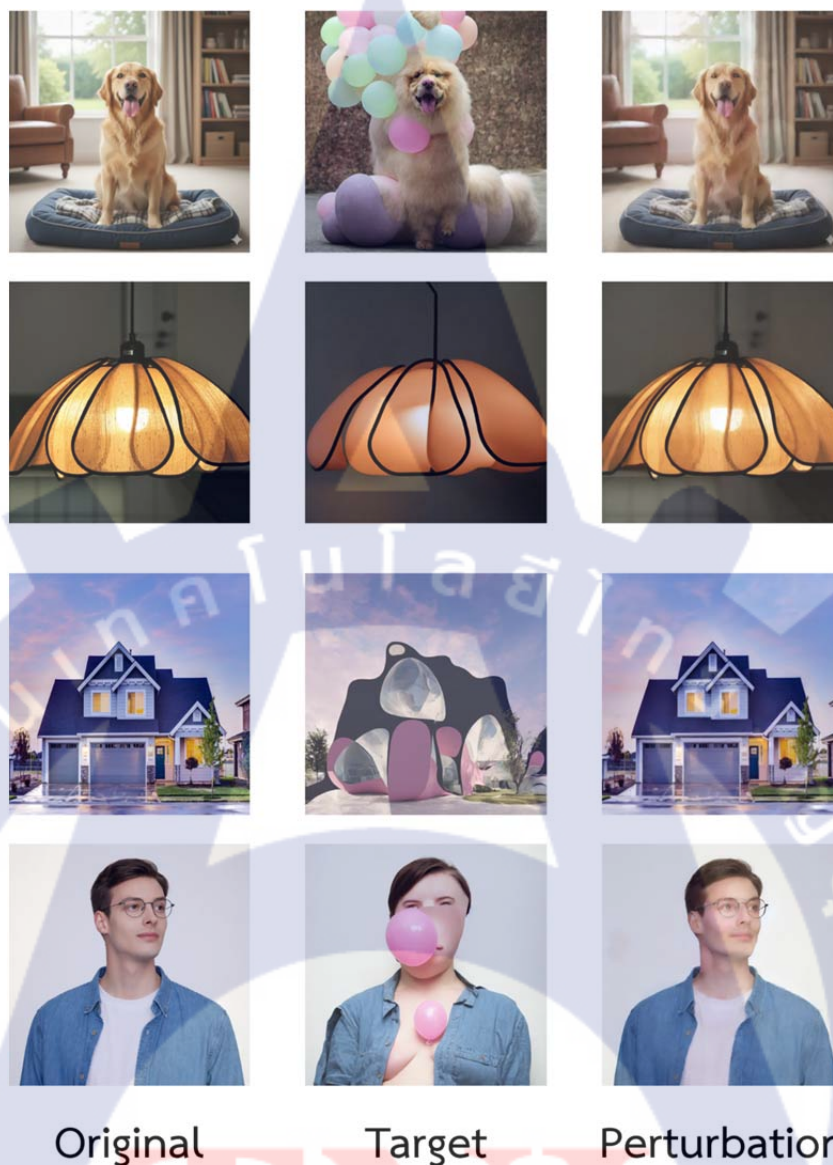
3.4.1 การเตรียมข้อมูลทดสอบ

ผู้วิจัยเลือกใช้ชุดข้อมูลภาพที่ประกอบด้วยภาพต้นฉบับ (Original Image) และภาพเป้าหมาย (Target Image) ซึ่งเป็นภาพต้นฉบับที่ถูกดัดแปลงให้เกิดความผิดปกติด้วยโมเดล Stable Diffusion 1.5 โดยชุดข้อมูลถูกจัดหมวดหมู่เป็นหลายประเภท ได้แก่ สัตว์ สิ่งของ สถานที่ และบุคคล รวมจำนวน 20 ภาพ เพื่อทดสอบประสิทธิภาพของวิธีการกับภาพที่มีลักษณะและองค์ประกอบแตกต่างกัน

ภาพเป้าหมายถูกสร้างด้วยกระบวนการ Image-to-Image ของโมเดล Stable Diffusion 1.5 โดยกำหนดค่า Random Seed เท่ากับ 123456 เพื่อให้สามารถสร้างซ้ำได้ โดย Prompt ที่ใช้มีโครงสร้างหลักเหมือนกันทุกภาพ และเปลี่ยนเฉพาะคำระบุชนิดของวัตถุ (เช่น Dog, Cat, Glass, Home, Person เป็นต้น) เท่านั้น เพื่อควบคุมรูปแบบเชิงภาพให้ใกล้เคียงกันในทุกกรณี

Prompt ดังกล่าวกำหนดลักษณะความผิดปกติในทิศทางเดียวกัน ได้แก่ โทนสี Pastel Pink รูปร่างพองกลม (Inflated Round Body หรือ Inflated Rounded Shape) โครงสร้างคล้ายลูกโป่ง (Balloon-Like Torso หรือ Balloon-Like Walls) ลักษณะนุ่มฟู (Soft Puffy Form) และสัดส่วนที่ผิดเพี้ยน (Distorted Proportions) เพื่อให้ภาพเป้าหมายมีรูปแบบความบิดเบือนที่สอดคล้องกันทุกหมวดหมู่

เพื่อแสดงลักษณะของข้อมูลที่ใช้ในการทดลอง ผู้วิจัยได้ยกตัวอย่างภาพต้นฉบับ (Original Image) ภาพเป้าหมาย (Target Image) และภาพที่ผ่านการบดบัง (Perturbation Image) ดังแสดงในรูปที่ 3.3 ซึ่งประกอบด้วยภาพจากหลายหมวดหมู่ ได้แก่ สัตว์ สิ่งของ สถานที่ และบุคคล



รูปที่ 3.3 ตัวอย่างภาพต้นฉบับ (Original Image) ภาพเป้าหมาย (Target Image) และภาพที่ผ่านการป้องกัน (Perturbation Image) ในแต่ละหมวดหมู่ ได้แก่ สัตว์ สิ่งของ สถานที่ และบุคคล

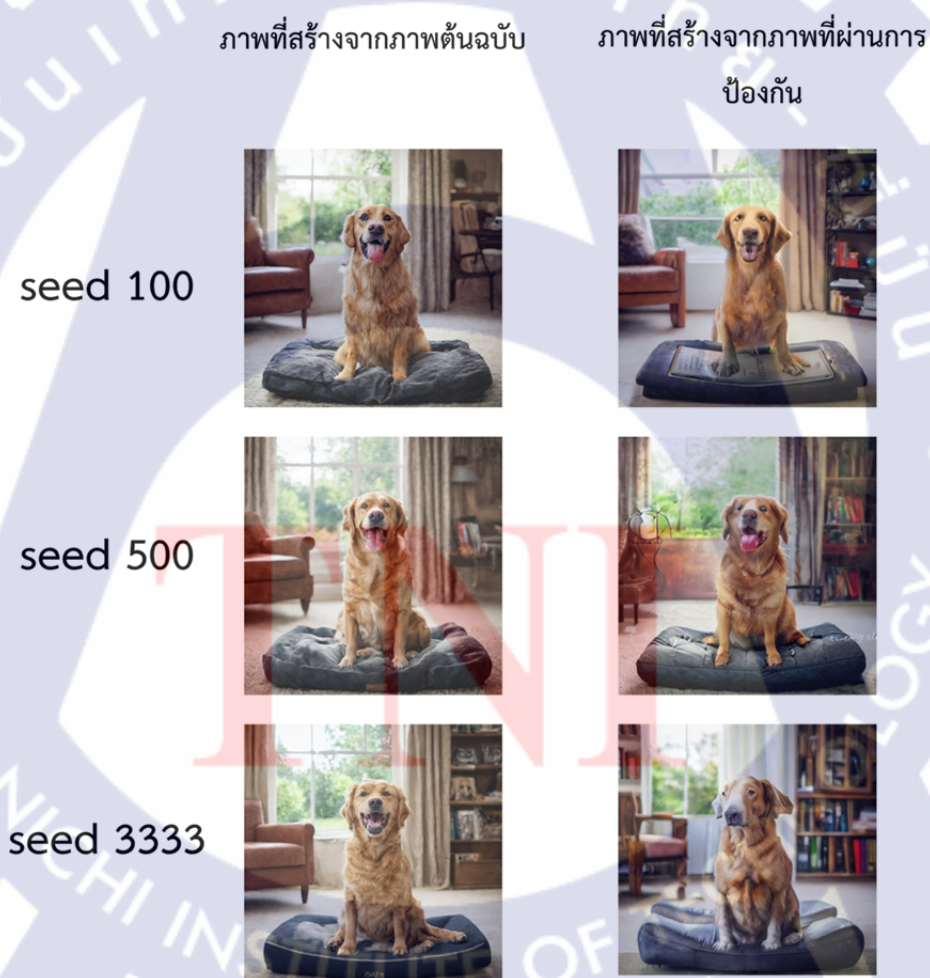
3.4.2 การสร้างการรบกวนเชิงแฝง (Perturbation, δ)

ทำการคำนวณหาค่าการรบกวนเชิงแฝง (δ) ที่เหมาะสมสำหรับภาพแต่ละคู่ (Original-Target) โดยใช้กระบวนการเพิ่มประสิทธิภาพเชิงเกรเดียนต์ (Gradient-based Optimization) ภายใต้ฟังก์ชันความสูญเสียหลัก จากนั้นนำค่าการรบกวนที่ได้มาฝังลงใน latent space ของภาพต้นฉบับ และถอดรหัสกลับมาเป็นภาพใหม่ที่เรียกว่า ภาพที่ผ่านการป้องกัน (Perturbation Image) ซึ่งมีลักษณะภายนอกใกล้เคียงกับต้นฉบับแต่แฝงการรบกวนไว้ภายใน

3.4.3 การทดลองด้วยโมเดล Stable Diffusion 1.5 (Image-to-Image Generation)

นำภาพต้นฉบับ (Original Image) และภาพที่ผ่านการป้องกัน (Perturbation Image) เข้าสู่โมเดล Stable Diffusion 1.5 ภายใต้โหมดการทำงานแบบ Image-to-Image (img2img) โดยกำหนดค่าพารามิเตอร์ให้เหมือนกันทุกประการ และใช้ Prompt เดียวกันคือ “A Realistic Photo of a ..., Natural Colors, Balanced Exposure”

ทั้งนี้ กระบวนการ Diffusion เป็นกระบวนการเชิงสุ่มที่ขึ้นกับค่า Random Seed ส่งผลให้ภาพที่สร้างขึ้นอาจแตกต่างกันได้ แม้จะใช้ภาพต้นฉบับและ Prompt เดียวกันก็ตาม ดังนั้น ผู้วิจัยจึงนำเสนอผลลัพธ์จากค่า Seed ที่แตกต่างกัน เพื่อแสดงความแปรผันของภาพ ดังแสดงในรูปที่ 3.3 ซึ่งเป็นการเปรียบเทียบภาพที่สร้างจากภาพต้นฉบับและภาพที่สร้างจากภาพที่ผ่านการป้องกันภายใต้ค่า Seed ต่างๆ



รูปที่ 3.4 ตัวอย่างภาพที่สร้างจากภาพต้นฉบับและภาพที่ผ่านการป้องกันภายใต้ค่า Seed ที่แตกต่างกัน

อย่างไรก็ตาม ในการทดลองหลักนี้ได้กำหนดใช้ค่า Seed เท่ากับ 3333 ในทุกกรณี เพื่อควบคุมตัวแปรและให้สามารถเปรียบเทียบผลลัพธ์ระหว่างภาพต้นฉบับและภาพที่ผ่านการป้องกันได้อย่างเป็นธรรม ทั้งนี้ เนื่องจากกระบวนการ Diffusion มีลักษณะเชิงสุ่ม (Stochastic Process) ผลลัพธ์ที่ได้อาจแตกต่างกันแม้ใช้ภาพและ Prompt เดียวกัน ดังนั้น การกำหนดค่า Seed คงที่จึงมีความสำคัญต่อความสามารถในการทำซ้ำของการทดลอง (Reproducibility) และช่วยให้ความแตกต่างที่สังเกตได้เกิดจากผลของ Perturbation อย่างแท้จริง ไม่ใช่จากความแปรผันเชิงสุ่มของกระบวนการสร้างภาพ ซึ่งผลลัพธ์ที่ได้จากกระบวนการนี้ประกอบด้วย

- ภาพที่สร้างจากภาพต้นฉบับ (Ori_img2img)
- ภาพที่สร้างจากภาพที่ผ่านการป้องกัน (Pert_img2img)

3.4.4 การเปรียบเทียบผลลัพธ์และการประเมินผล

ผลลัพธ์จากกระบวนการทดลองถูกนำมาวิเคราะห์ในสองส่วน ได้แก่ การประเมินผลกระทบของ Perturbation ต่อภาพต้นฉบับ และการประเมินผลกระทบต่อพฤติกรรมการสร้างภาพของโมเดล Stable Diffusion 1.5

ในส่วนแรก ผู้วิจัยทำการเปรียบเทียบระหว่างภาพต้นฉบับ (Original Image) และภาพที่ผ่านการป้องกัน (Perturbation Image) เพื่อประเมินว่าการฝังการรบกวนเชิงแฝงส่งผลต่อคุณภาพและความหมายของภาพต้นฉบับมากน้อยเพียงใด โดยมุ่งตรวจสอบว่าภาพที่ผ่านการป้องกันยังคงรักษาโครงสร้างและความหมายหลักของวัตถุไว้ได้หรือไม่

ในส่วนถัดมา ผู้วิจัยทำการเปรียบเทียบระหว่างภาพที่สร้างจากภาพต้นฉบับ (Ori_img2img) และภาพที่สร้างจากภาพที่ผ่านการป้องกัน (Pert_img2img) เพื่อวิเคราะห์ว่าการฝัง Perturbation ส่งผลต่อกระบวนการ Diffusion ของโมเดลอย่างไร และสามารถทำให้ภาพที่สร้างขึ้นเกิดความเบี่ยงเบนจากผลลัพธ์ปกติได้มากน้อยเพียงใด

การประเมินผลทั้งสองกรณีใช้ตัวชี้วัดเชิงปริมาณชุดเดียวกัน ได้แก่ SSIM, PSNR, CLIP Similarity และ LPIPS เพื่อให้สามารถวิเคราะห์ผลลัพธ์ได้อย่างเป็นระบบทั้งในเชิงโครงสร้าง เชิงการรับรู้ และเชิงความหมาย โดยค่าที่ได้สะท้อนให้เห็นทั้งระดับความใกล้เคียงของภาพและระดับความเบี่ยงเบนที่เกิดจากการรบกวน

3.5 การประเมินผลการวิจัย

การประเมินผลการทดลองในงานวิจัยนี้ใช้ตัวชี้วัด 4 ประเภท ได้แก่ SSIM, PSNR, CLIP Similarity และ LPIPS ตามที่อธิบายไว้ในบทที่ 2 โดยนำมาใช้เพื่อวิเคราะห์ความแตกต่างของภาพทั้งในเชิงโครงสร้าง เชิงพิกเซล เชิงความหมาย และเชิงการรับรู้ของมนุษย์ รายละเอียดการใช้งานของตัวชี้วัดแต่ละประเภทมีดังนี้

1) SSIM (Structural Similarity Index Measure)

ใช้ประเมินระดับความคล้ายคลึงเชิงโครงสร้าง ความสว่าง และความแตกต่างของความเข้มแสงระหว่างภาพสองภาพ โดยค่า SSIM ที่สูงบ่งชี้ว่าภาพทั้งสองยังคงมีโครงสร้างใกล้เคียงกัน ขณะที่ค่าที่ต่ำสะท้อนถึงการเปลี่ยนแปลงของโครงสร้างภาพที่เพิ่มขึ้น

2) PSNR (Peak Signal-to-Noise Ratio)

ใช้วัดระดับความแตกต่างของค่าพิกเซลระหว่างภาพสองภาพ โดยค่า PSNR ที่สูงหมายความว่าความแตกต่างเชิงพิกเซลอยู่ในระดับต่ำ ในขณะที่ค่า PSNR ที่ต่ำแสดงว่ารายละเอียดของภาพมีการเปลี่ยนแปลงเพิ่มขึ้น

3) CLIP Similarity

ใช้ประเมินความสอดคล้องเชิงความหมาย (Semantic Similarity) ระหว่างภาพสองภาพผ่านโมเดล CLIP โดยค่าที่สูงแสดงว่าภาพยังคงรักษาความหมายหลักของวัตถุไว้ได้ ขณะที่ค่าที่ลดลงสะท้อนถึงการเบี่ยงเบนเชิงความหมายของภาพภายใต้กระบวนการสร้างภาพ

4) LPIPS (Learned Perceptual Image Patch Similarity)

ใช้วัดความแตกต่างเชิงการรับรู้ (Perceptual Difference) ระหว่างภาพสองภาพ โดยค่า LPIPS ที่สูงขึ้นบ่งชี้ว่าภาพมีความแตกต่างกันมากขึ้นในมุมมองของมนุษย์

นอกจากการอธิบายตัวชี้วัด ผู้วิจัยได้กำหนดเกณฑ์การตีความผลลัพธ์โดยแบ่งเป็น 2 กรณี ได้แก่ การประเมินคุณภาพของภาพต้นฉบับและภาพที่ผ่านการป้องกัน และการประเมินผลกระทบของภาพที่ผ่านการป้องกันต่อกระบวนการสร้างภาพของโมเดล โดยในกรณีแรก ค่าที่ดีควรแสดงถึงความใกล้เคียงกันของภาพ (SSIM, PSNR, CLIP สูง และ LPIPS ต่ำ) เพื่อสะท้อนการรักษาคุณภาพของภาพต้นฉบับ ขณะที่กรณีที่สอง ค่าที่ดีควรแสดงถึงความแตกต่างของภาพ (SSIM, PSNR, CLIP ลดลง และ LPIPS สูงขึ้น) เพื่อสะท้อนผลของ Perturbation ที่มีต่อพฤติกรรมของโมเดล

บทที่ 4

ผลการวิจัย

บทนี้นำเสนอผลการทดลองของวิธีการฝังการรบกวน (Perturbation) ใน Latent Space เพื่อป้องกันภาพจากการถูกนำไปใช้งานโดยโมเดล Stable Diffusion 1.5 ซึ่งจะใช้ตัวชี้วัดชุดเดียวกัน ได้แก่ SSIM, PSNR, CLIP Similarity และ LPIPS เพื่อให้สามารถเปรียบเทียบผลลัพธ์ได้อย่างเป็นธรรม และสอดคล้องกัน โดยแบ่งการประเมินออกเป็น 2 ส่วนหลักได้แก่

4.1 การประเมินคุณภาพของภาพต้นฉบับและภาพที่ผ่านการป้องกัน

4.2 การประเมินผลกระทบของภาพที่ผ่านการป้องกันต่อกระบวนการสร้างภาพของโมเดล

4.1 การประเมินคุณภาพของภาพต้นฉบับและภาพที่ผ่านการป้องกัน

การทดลองในส่วนนี้เป็นการเปรียบเทียบระหว่างภาพต้นฉบับ (Original Image) และภาพที่ผ่านการป้องกัน (Perturbation Image) โดยใช้ตัวชี้วัดชุดเดียวกันที่กล่าวไว้ข้างต้น เพื่อประเมินความแตกต่างเชิงโครงสร้าง เชิงพิกเซล เชิงความหมาย และเชิงการรับรู้ของมนุษย์

เพื่อแสดงผลการประเมินเชิงตัวเลขของคุณภาพภาพต้นฉบับและภาพที่ผ่านการป้องกันดังตารางที่ 4.1 ค่าตัวชี้วัดของภาพแต่ละกลุ่มตัวอย่าง

ตารางที่ 4.1 ค่าตัวชี้วัด SSIM, PSNR, CLIP Similarity และ LPIPS ของภาพต้นฉบับและภาพที่ผ่านการป้องกัน

Image	SSIM	PSNR	Clip	LPIPS
Dog	0.8608	27.80	0.9624	0.1579
Cat	0.8874	27.95	0.9599	0.1485
Rabbit	0.7203	25.18	0.9882	0.2060
Parrot	0.8128	25.22	0.9669	0.1782
Butterfly	0.8434	27.43	0.9916	0.1519
Glass	0.8429	27.38	0.9675	0.1657
Lamp	0.6717	26.59	0.9590	0.3338
Phone	0.8055	27.52	0.9694	0.2369
Pillow	0.7888	27.52	0.9356	0.2468
Balloon	0.8454	27.49	0.9822	0.0978

ตารางที่ 4.1 ค่าตัวชี้วัด SSIM, PSNR, CLIP Similarity และ LPIPS ของภาพต้นฉบับและภาพที่ผ่านการป้องกัน (ต่อ)

Image	SSIM	PSNR	Clip	LPIPS
Home	0.8063	25.87	0.9858	0.1456
Garden	0.4228	19.95	0.8984	0.4229
Beach	0.7921	28.42	0.9657	0.1986
Parking	0.6276	23.80	0.9523	0.2729
Street	0.7350	26.23	0.9351	0.2577
Person1	0.8548	27.49	0.9455	0.1502
Person2	0.8467	27.99	0.9036	0.1294
Person3	0.8340	27.36	0.9734	0.1329
Person4	0.6729	27.40	0.9153	0.3040
Person5	0.8239	27.27	0.9474	0.1241

ผลการทดลองพบว่า ค่า SSIM ของคู่ Original–Perturbation อยู่ในระดับค่อนข้างสูง แสดงให้เห็นว่าภาพที่ผ่านการป้องกันยังคงโครงสร้างหลักของภาพต้นฉบับไว้ได้ ค่า PSNR อยู่ในช่วงประมาณ 25-28 dB สะท้อนว่าความแตกต่างเชิงพิกเซลอยู่ในระดับที่ไม่ทำให้คุณภาพภาพลดลงอย่างชัดเจน

ในด้านความหมาย ค่า CLIP Similarity ยังคงอยู่ในระดับสูง (มากกว่า 0.94 โดยเฉลี่ย) แสดงว่าภาพที่ผ่านการป้องกันยังคงถูกตีความว่าเป็นวัตถุประเภทเดิม เช่น สุนัข กระต่าย หรือบุคคล ขณะที่ค่า LPIPS อยู่ในระดับต่ำถึงปานกลาง (ประมาณ 0.10-0.30) บ่งชี้ว่าความแตกต่างเชิงการรับรู้ของมนุษย์ (Perceptual) อยู่ในระดับที่มนุษย์แทบไม่สังเกตเห็น

จากค่าเฉลี่ยของตัวชี้วัดพบว่า SSIM มีค่าเฉลี่ยเท่ากับ 0.7748 และ PSNR เท่ากับ 26.59 dB แสดงให้เห็นว่าภาพที่ผ่านการป้องกันยังคงรักษาโครงสร้างและคุณภาพเชิงพิกเซลของภาพต้นฉบับไว้ได้ในระดับที่ยอมรับได้ ขณะที่ค่า CLIP Similarity มีค่าเฉลี่ยสูงถึง 0.9553 สะท้อนว่าภาพยังคงถูกตีความว่าเป็นวัตถุประเภทเดิมในเชิงความหมาย ส่วนค่า LPIPS มีค่าเฉลี่ยอยู่ที่ 0.2031 ซึ่งอยู่ในระดับต่ำถึงปานกลาง บ่งชี้ว่าความแตกต่างเชิงการรับรู้ของมนุษย์ (Perceptual) หมายถึง ความแตกต่างที่มนุษย์สามารถมองเห็นได้) อยู่ในระดับที่ไม่เด่นชัด

ผลลัพธ์ดังกล่าวชี้ให้เห็นว่า Perturbation ที่นำเสนอสามารถฝังลงในภาพได้โดยไม่ส่งผลกระทบต่อคุณภาพและความหมายของภาพในมุมมองของมนุษย์

เพื่อยืนยันผลลัพธ์จากค่าตัวชี้วัดในตารางที่ 4.1 ร่วมกับการพิจารณาเชิงภาพ รูปที่ 4.1-4.4 แสดงตัวอย่างการเปรียบเทียบระหว่างภาพต้นฉบับและภาพที่ผ่านการป้องกันในกลุ่มตัวอย่างประเภท สัตว์ สิ่งของ สถานที่ และบุคคล



รูปที่ 4.1 การเปรียบเทียบระหว่างภาพต้นฉบับและภาพที่ผ่านการป้องกัน ในกลุ่มตัวอย่างประเภทสัตว์

ภาพต้นฉบับ

ภาพที่ผ่านการป้องกัน

Glass



Lamp



Phone



Pillow



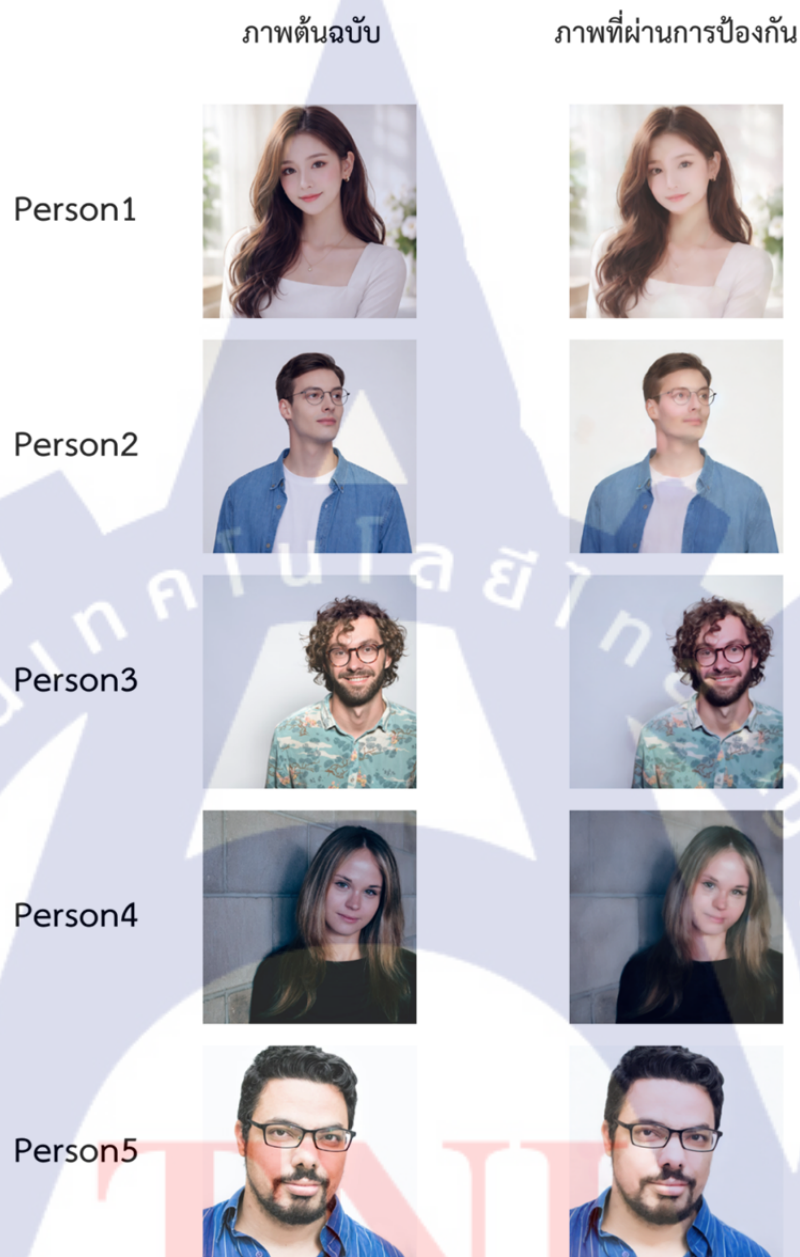
Balloon



รูปที่ 4.2 การเปรียบเทียบระหว่างภาพต้นฉบับและภาพที่ผ่านการป้องกัน ในกลุ่มตัวอย่างประเภทสิ่งของ



รูปที่ 4.3 การเปรียบเทียบระหว่างภาพต้นฉบับและภาพที่ผ่านการป้องกัน ในกลุ่มตัวอย่างประเภทสถานที่



รูปที่ 4.4 การเปรียบเทียบระหว่างภาพต้นฉบับและภาพที่ผ่านการป้องกัน ในกลุ่มตัวอย่างประเภทบุคคล

4.2 การประเมินผลกระทบของภาพที่ผ่านการป้องกันต่อกระบวนการสร้างภาพของโมเดล

ในการประเมินผลกระทบของ Perturbation ต่อพฤติกรรมการสร้างภาพของโมเดล Stable Diffusion 1.5 งานวิจัยนี้ได้นำภาพต้นฉบับ (Original Image) และภาพที่ผ่านการป้องกัน (Perturbation Image) มาใช้เป็นอินพุตในกระบวนการ Image-to-Image Generation (img2img) ภายใต้ Prompt และค่าพารามิเตอร์เดียวกันทุกประการ เพื่อควบคุมตัวแปรในการทดลองให้คงที่

จากนั้นนำภาพที่ได้จากภาพต้นฉบับ (Ori_img2img) และภาพที่ได้จากภาพที่ผ่านการป้องกัน (Pert_img2img) มาเปรียบเทียบเชิงปริมาณ โดยใช้ตัวชี้วัด SSIM, PSNR, CLIP Similarity และ LPIPS เพื่อวิเคราะห์ความแตกต่างทั้งในเชิงโครงสร้างภาพเชิง Perceptual และเชิงความหมาย (Semantic)

ตารางที่ 4.2 ค่าตัวชี้วัด SSIM, PSNR, CLIP Similarity และ LPIPS ของภาพต้นฉบับและภาพที่ผ่านการป้องกันที่สร้างจากกระบวนการ img2img ของ Stable Diffusion 1.5

Image	SSIM	PSNR	Clip	LPIPS
Dog	0.5056	15.91	0.8227	0.5365
Cat	0.6859	16.61	0.8799	0.4741
Rabbit	0.3487	15.97	0.9293	0.4858
Parrot	0.6306	16.03	0.9163	0.3688
Butterfly	0.4647	14.50	0.8878	0.5770
Glass	0.6307	16.04	0.8902	0.5012
Lamp	0.4893	16.94	0.8007	0.6466
Phone	0.6230	18.04	0.9248	0.5045
Pillow	0.6319	19.36	0.8910	0.5502
Balloon	0.7333	19.80	0.9118	0.2717
Home	0.5930	15.97	0.9230	0.3862
Garden	0.0936	11.83	0.7101	0.6408
Beach	0.7016	22.50	0.8314	0.3788
Parking	0.2503	13.84	0.8435	0.5665
Street	0.4546	16.63	0.7281	0.5673
Person1	0.6497	18.09	0.6566	0.4493
Person2	0.6836	19.13	0.4389	0.4558
Person3	0.6714	18.26	0.7627	0.3183
Person4	0.5035	19.88	0.8532	0.4718
Person5	0.6531	17.52	0.7045	0.3684

จากค่าเฉลี่ยของตัวชี้วัดพบว่า SSIM มีค่าเฉลี่ยเท่ากับ 0.5499 และ PSNR เท่ากับ 17.14 dB สะท้อนว่าภาพที่สร้างจากภาพต้นฉบับ (Ori_img2img) และภาพที่สร้างจากภาพที่ผ่านการป้องกัน (Pert_img2img) ภายใต้กระบวนการ Image-to-Image มีความแตกต่างเชิงโครงสร้างและเชิงพิกเซลในระดับค่อนข้างสูง

ในด้านความหมาย ค่า CLIP Similarity มีค่าเฉลี่ยเท่ากับ 0.8153 แสดงว่าโมเดลยังสามารถรักษาความหมายหลักของวัตถุไว้ได้บางส่วน อย่างไรก็ตาม ระดับความสอดคล้องเชิงความหมายระหว่างภาพทั้งสองลดลงในระดับที่สังเกตได้

ขณะเดียวกัน ค่า LPIPS มีค่าเฉลี่ยเท่ากับ 0.4760 ซึ่งอยู่ในระดับปานกลางถึงค่อนข้างสูง บ่งชี้ว่าภาพทั้งสองมีความแตกต่างเชิงการรับรู้ของมนุษย์ (Perceptual Difference) อย่างชัดเจน

ผลลัพธ์ดังกล่าวสะท้อนว่า Perturbation ที่ฝังใน Latent Space สามารถส่งผลต่อพฤติกรรมการสร้างภาพของโมเดล Stable Diffusion 1.5 ได้อย่างมีนัยสำคัญ โดยเฉพาะเมื่อภาพถูกนำเข้าสู่กระบวนการ Image-to-Image ภายใต้ Prompt และพารามิเตอร์เดียวกัน

เพื่อแสดงให้เห็นผลกระทบของ Perturbation ในเชิงคุณภาพ รูปที่ 4.5-4.8 แสดงตัวอย่างการเปรียบเทียบภาพที่สร้างจากภาพต้นฉบับและภาพที่สร้างจากภาพที่ผ่านการป้องกัน โดยใช้กระบวนการ Image-to-Image ภายใต้ Prompt และพารามิเตอร์เดียวกัน

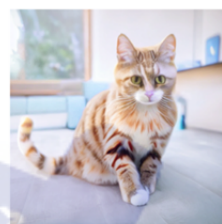
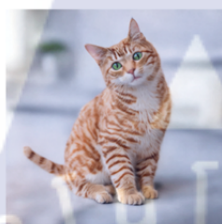
ภาพที่สร้างจากภาพต้นฉบับ

ภาพที่สร้างจากภาพที่ผ่านการ
ป้องกัน

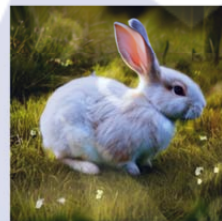
Dog



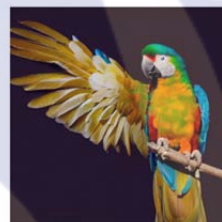
Cat



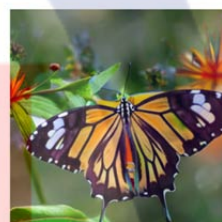
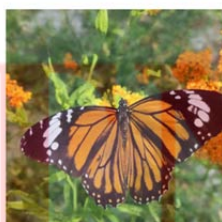
Rabbit



Parrot



Butterfly



รูปที่ 4.5 ตัวอย่างผลลัพธ์ของ Stable Diffusion 1.5 ของภาพต้นฉบับ (Original Image) และภาพที่ผ่านการป้องกัน (Perturbation Image) ประเภทสัตว์

ภาพที่สร้างจากภาพต้นฉบับ

ภาพที่สร้างจากภาพที่ผ่านการ
ป้องกัน

Glass



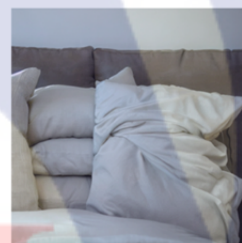
Lamp



Phone



Pillow



Balloon



รูปที่ 4.6 ตัวอย่างผลลัพธ์ของ Stable Diffusion 1.5 ของภาพต้นฉบับ (Original Image) และภาพที่ผ่านการป้องกัน (Perturbation Image) ประเภทสิ่งของ

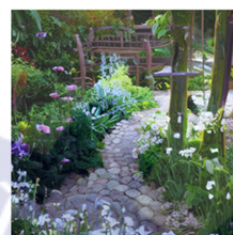
ภาพที่สร้างจากภาพต้นฉบับ

ภาพที่สร้างจากภาพที่ผ่านการ
ป้องกัน

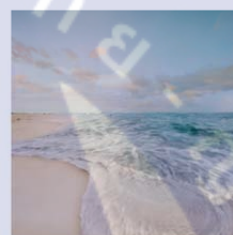
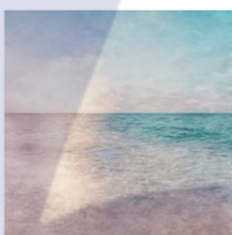
Home



Garden



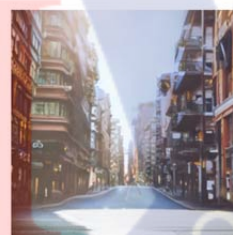
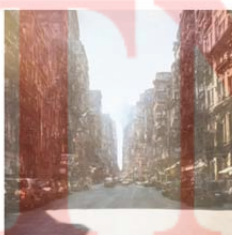
Beach



Parking



Street



รูปที่ 4.7 ตัวอย่างผลลัพธ์ของ Stable Diffusion 1.5 ของภาพต้นฉบับ (Original Image) และภาพที่ผ่านการป้องกัน (Perturbation Image) ประเภทสถานที่

ภาพที่สร้างจากภาพต้นฉบับ

ภาพที่สร้างจากภาพที่ผ่านการ
ป้องกัน

Person1



Person2



Person3



Person4



Person5



รูปที่ 4.8 ตัวอย่างผลลัพธ์ของ Stable Diffusion 1.5 ของภาพต้นฉบับ (Original Image) และภาพที่ผ่านการป้องกัน (Perturbation Image) ประเภทบุคคล

บทที่ 5

สรุปผลและข้อเสนอแนะ

จากการวิจัยเรื่อง การออกแบบและพัฒนาวิธีการป้องกันภาพที่มีลิขสิทธิ์จากการถูกนำไปใช้งานในระบบ Generative AI ด้วยการฝัง Perturbation ใน Latent Space มีผลสรุปดังนี้

1. สรุปผลการวิจัย
2. อภิปรายผลการวิจัย
3. ปัญหาและข้อจำกัด
4. ข้อเสนอแนะ

5.1 สรุปผลการวิจัย

งานวิจัยนี้มีวัตถุประสงค์เพื่อพัฒนาวิธีการป้องกันภาพที่มีลิขสิทธิ์จากการถูกนำไปใช้งานโดยโมเดล Stable Diffusion 1.5 ผ่านการฝัง Perturbation ใน Latent Space โดยมุ่งเน้นให้ภาพที่ผ่านการป้องกันยังคงมีลักษณะใกล้เคียงกับภาพต้นฉบับในมุมมองของมนุษย์ ขณะเดียวกันสามารถส่งผลกระทบต่อพฤติกรรมการสร้างภาพของโมเดลภายใต้กระบวนการ Image-to-Image ได้อย่างมีนัยสำคัญ

การประเมินผลประกอบด้วยสองส่วนหลัก ได้แก่ การเปรียบเทียบระหว่างภาพต้นฉบับและภาพที่ผ่านการป้องกัน เพื่อประเมินคุณภาพของภาพที่ถูกฝัง Perturbation และการเปรียบเทียบระหว่างภาพที่สร้างจากภาพต้นฉบับและภาพที่สร้างจากภาพที่ผ่านการป้องกันผ่านกระบวนการ Image-to-Image ภายใต้ Prompt และพารามิเตอร์เดียวกัน เพื่อประเมินผลกระทบของ Perturbation ต่อพฤติกรรมการสร้างภาพของโมเดล โดยใช้ตัวชี้วัด SSIM, PSNR, CLIP Similarity และ LPIPS

ผลการทดลองในส่วนแรกพบว่า ภาพที่ผ่านการป้องกันยังคงรักษาโครงสร้างและคุณภาพเชิงพิกเซลของภาพต้นฉบับไว้ได้ในระดับที่ยอมรับได้ โดยมีค่า SSIM และ PSNR อยู่ในระดับค่อนข้างสูง (โดยมีค่าเฉลี่ย SSIM เท่ากับ 0.7748 และค่า PSNR เท่ากับ 26.59 dB) ขณะที่ค่า CLIP Similarity ยังคงอยู่ในระดับสูง (CLIP = 0.9553) และค่า LPIPS อยู่ในระดับต่ำถึงปานกลาง (LPIPS = 0.2031) สะท้อนว่า Perturbation สามารถฝังลงในภาพได้โดยไม่ส่งผลกระทบต่อคุณภาพและความหมายของภาพในมุมมองของมนุษย์

ในส่วนที่สอง ซึ่งเป็นการประเมินระหว่างภาพที่สร้างจากภาพต้นฉบับและภาพที่สร้างจากภาพที่ผ่านการป้องกันผ่านกระบวนการ Image-to-Image พบว่าค่าตัวชี้วัด SSIM และ PSNR อยู่ในระดับต่ำ (SSIM = 0.5499 และ PSNR = 17.14 dB) ขณะที่ค่า LPIPS อยู่ในระดับปานกลางถึงค่อนข้างสูง (LPIPS = 0.4760) และค่า CLIP Similarity อยู่ในระดับต่ำ (CLIP = 0.8153) สะท้อนว่าภาพผลลัพธ์

ทั้งสองมีความแตกต่างกันอย่างชัดเจนทั้งในเชิงโครงสร้าง เชิงการรับรู้ และเชิงความหมาย ภายใต้เงื่อนไขการสร้างภาพเดียวกัน ความแตกต่างดังกล่าวบ่งชี้ว่า perturbation สามารถเปลี่ยนทิศทางของกระบวนการ Diffusion และรบกวน Latent Representation ของโมเดลได้อย่างมีประสิทธิภาพ

นอกจากนี้ จากการวิเคราะห์เชิงคุณภาพพบว่า Perturbation ส่งผลกระทบต่อภาพกลุ่มสัตว์และบุคคลได้เด่นชัด โดยโมเดลมักสร้างภาพที่เกิดการเบี่ยงเบนเชิงความหมาย เช่น การเปลี่ยนลักษณะใบหน้า รูปร่าง หรืออัตลักษณ์ของวัตถุ ขณะที่ภาพประเภทสิ่งของและสถานที่บางกลุ่มยังคงรักษาโครงสร้างหลักไว้ได้ในระดับหนึ่ง ซึ่งสะท้อนถึงความแตกต่างของลักษณะ Latent Representation ตามประเภทของวัตถุ

โดยสรุป วิธีการฝัง Perturbation ใน Latent Space ที่นำเสนอสามารถรักษาคุณภาพของภาพที่ผ่านการป้องกันในระดับที่มนุษย์ยอมรับได้ และในขณะเดียวกันสามารถสร้างความแตกต่างของผลลัพธ์การสร้างภาพเมื่อถูกนำไปใช้กับ Stable Diffusion 1.5 ภายใต้กระบวนการ Image-to-Image ได้อย่างมีนัยสำคัญ แสดงให้เห็นถึงศักยภาพของแนวทางนี้ในการใช้เป็นกลไกป้องกันภาพที่มีลิขสิทธิ์จากการถูกนำไปใช้งานในระบบ Generative AI

5.2 อภิปรายผลการวิจัย

จากผลการทดลองสามารถอภิปรายได้ว่า วิธีการฝัง Perturbation ใน Latent Space สามารถรักษาคุณภาพของภาพต้นฉบับไว้ได้ในระดับที่มนุษย์ยอมรับได้ ขณะเดียวกันสามารถรบกวนพฤติกรรมของโมเดลในกระบวนการ Image-to-Image ให้เกิดผลลัพธ์ที่แตกต่างอย่างชัดเจน ผลลัพธ์ดังกล่าวสะท้อนว่า Perturbation ที่ฝังใน Latent Space ส่งผลต่อการแทนข้อมูลภายในโมเดลโดยตรง ทำให้ทิศทางของกระบวนการ Diffusion เปลี่ยนแปลง เมื่อเปรียบเทียบกับงานวิจัยที่เกี่ยวข้อง พบว่าวิธีการนี้มีแนวคิดสอดคล้องกับ PhotoGuard และ DiffusionShield ที่ใช้การรบกวนข้อมูลเพื่อป้องกันการใช้งานโดย Generative AI แต่มีจุดเด่นที่การดำเนินการใน Latent Space ซึ่งช่วยลดผลกระทบต่อคุณภาพของภาพ

ในด้านการใช้งาน วิธีการนี้สามารถนำไปใช้ปกป้องลิขสิทธิ์ภาพในแพลตฟอร์มดิจิทัล และช่วยลดความเสี่ยงของการนำภาพไปใช้โดยไม่ได้รับอนุญาต อย่างไรก็ตาม ประสิทธิภาพของวิธีการยังขึ้นอยู่กับประเภทของภาพ ซึ่งควรมีการศึกษาเพิ่มเติมในอนาคต

5.3 ปัญหาและข้อจำกัด

แม้ว่าวิธีการฝัง Perturbation ใน Latent Space ที่นำเสนอในงานวิจัยนี้จะสามารถรักษาคุณภาพของภาพต้นฉบับไว้ได้ในระดับที่ยอมรับได้ และสามารถสร้างความแตกต่างของผลลัพธ์ภาพในกระบวนการ Image-to-Image ได้อย่างมีนัยสำคัญ แต่จากผลการทดลองยังพบข้อจำกัดหลายประการที่ควรพิจารณา

ในบางกรณี การฝัง Perturbation ส่งผลให้รายละเอียดของภาพต้นฉบับลดลง หรือเกิดความผิดปกติทางรูปลักษณะบางส่วน แม้ว่าค่าตัวชี้วัดเชิงปริมาณ เช่น SSIM และ PSNR จะยังอยู่ในระดับที่เหมาะสมก็ตาม สถานการณ์ดังกล่าวสะท้อนถึงความท้าทายในการควบคุมสมดุลระหว่างการรักษาคุณภาพของภาพ (Original-Perturbation) และการเพิ่มระดับความแตกต่างของผลลัพธ์ในกระบวนการ Image-to-Image (Ori_img2img-Pert_img2img)

นอกจากนี้ ประสิทธิภาพของ Perturbation มีความแตกต่างตามประเภทของภาพ โดยจากผลการทดลองพบว่า ภาพกลุ่มสัตว์และบุคคลมีแนวโน้มแสดงความแตกต่างของผลลัพธ์ในกระบวนการ image-to-image ได้ชัดเจนกว่า ขณะที่ภาพประเภทสิ่งปลูกสร้างและวัตถุทั่วไปยังคงรักษาโครงสร้างและความหมายหลักไว้ได้ในระดับหนึ่ง ภายใต้เงื่อนไข Prompt และพารามิเตอร์เดียวกัน ซึ่งสะท้อนถึงความแตกต่างของลักษณะการแทนข้อมูล (Latent Representation) ตามประเภทของวัตถุ

ขณะเดียวกัน กระบวนการสร้าง Perturbation ใน Latent Space จำเป็นต้องอาศัยการปรับค่าแบบวนซ้ำร่วมกับการคำนวณ Gradient และการเข้ารหัส-ถอดรหัสผ่าน VAE หลายรอบ ส่งผลให้การป้องกันภาพแต่ละภาพใช้ระยะเวลาในการประมวลผลค่อนข้างสูงเมื่อเทียบกับการประมวลผลภาพทั่วไป ข้อจำกัดด้านเวลาและทรัพยากรการคำนวณดังกล่าวอาจส่งผลกระทบต่อความเหมาะสมในการนำวิธีการไปใช้งานกับภาพจำนวนมาก หรือในบริบทที่ต้องการการประมวลผลแบบรวดเร็ว

นอกจากนี้ งานวิจัยนี้จำกัดการทดลองเฉพาะโมเดล Stable Diffusion 1.5 และประเมินผลผ่านกระบวนการ Image-to-Image ภายใต้เงื่อนไขที่ควบคุมตัวแปรไว้คงที่ จึงยังไม่สามารถสรุปได้อย่างชัดเจนถึงความสามารถของ Perturbation เมื่อประยุกต์ใช้กับโมเดล Diffusion รุ่นอื่น หรือภายใต้เงื่อนไขการสร้างภาพที่แตกต่างออกไป

ข้อจำกัดทั้งหมดนี้สะท้อนว่าวิธีการที่น่าเสนอแม้จะแสดงศักยภาพในการป้องกันภาพจากการถูกนำไปใช้ในกระบวนการสร้างภาพ แต่ยังมีพื้นที่สำหรับการปรับปรุงในด้านความสม่ำเสมอของผลลัพธ์ ประสิทธิภาพเชิงการคำนวณ และความสามารถในการประยุกต์ใช้ในบริบทที่หลากหลายยิ่งขึ้น

5.4 ข้อเสนอแนะ

จากผลการทดลองและข้อจำกัดที่พบ งานวิจัยในอนาคตสามารถขยายการศึกษาเพิ่มเติมในหลายแนวทาง เพื่อเพิ่มประสิทธิภาพและความสามารถในการประยุกต์ใช้งานของวิธีการที่น่าเสนอ

ในเบื้องต้น ควรทดสอบวิธีการฝัง Perturbation กับโมเดล Diffusion รุ่นใหม่หรือสถาปัตยกรรมที่แตกต่างออกไป เช่น Stable Diffusion XL เพื่อประเมินความสามารถในการถ่ายโอนผล (Transferability) ของ Perturbation ข้ามโมเดล และศึกษาความสม่ำเสมอของผลลัพธ์ในบริบทของสถาปัตยกรรมที่มีความซับซ้อนสูงขึ้น

นอกจากนี้ ควรศึกษาผลของ Perturbation ในบริบทของการฝึกโมเดล (Training-Time) หรือการ Fine-Tuning เพื่อประเมินว่าการฝัง Perturbation ใน Latent Space สามารถส่งผลกระทบต่อกระบวนการเรียนรู้ของโมเดลได้หรือไม่ ซึ่งจะช่วยเปิดมุมมองใหม่เกี่ยวกับการป้องกันภาพในระดับข้อมูลฝึก

อีกทั้งควรพัฒนา Perturbation ให้มีลักษณะปรับตัวตามประเภทของวัตถุ (Adaptive Perturbation) เนื่องจากผลการทดลองแสดงให้เห็นว่าประสิทธิภาพของวิธีมีความแตกต่างตามกลุ่มภาพ การออกแบบ Perturbation ที่คำนึงถึงลักษณะโครงสร้างของวัตถุแต่ละประเภทอาจช่วยเพิ่มความสม่ำเสมอของผลลัพธ์ และประสิทธิภาพในการป้องกัน

สุดท้ายนี้ควรศึกษาความทนทานของ Perturbation ต่อกระบวนการแปลงภาพ เช่น การบีบอัด การปรับขนาด หรือการปรับคุณภาพไฟล์ เพื่อให้แนวทางที่นำเสนอสามารถนำไปประยุกต์ใช้ได้ในสถานการณ์การใช้งาน



บรรณานุกรม

TNI

บรรณานุกรม

- [1] Extreme IT, “ศิลปินรวมตัวยื่นฟ้อง Stability AI และ Midjourney เหตุนำภาพวาดไปสอน AI โดยไม่ได้รับอนุญาต,” [Online]. Available : <https://www.extremeit.com/stable-diffusion-ai-midjourney/>. [Accessed : 29 กันยายน 2568].
- [2] Krungsri Research, “Generative AI เทคโนโลยีพลิกโฉมโลก,” [Online]. Available : <https://www.krungsri.com/th/research/research-intelligence/generative-ai-2023>. [Accessed : 29 กันยายน 2568].
- [3] Finnegan, “Getty Images v. Stability AI : The UK Court Battle that Could Reshape AI and Copyright Law,” [Online]. Available : <https://www.finnegan.com/en/insights/articles/getty-images-vs-stability-ai-the-uk-court-battle-that-could-reshape-ai-and-copyright-law.html>. [Accessed : September 29, 2025].
- [4] Loeb & Loeb LLP, “Andersen v. Stability AI Ltd.,” [Online]. Available : <https://www.loeb.com/en/insights/publications/2023/11/andersen-v-stability-ai-ltd>. [Accessed : September 29, 2025].
- [5] A. Bandi et al., “The power of generative AI : A review of requirements, models, input-output formats, evaluation metrics, and challenges,” *Future Internet*, vol. 15, no. 2, pp. 222, May-June 2023.
- [6] I. J. Goodfellow et al., “Explaining and harnessing adversarial examples,” *International Conference on Learning Representations, ICLR 2015*, San Diego, CA, USA, May 7-9, 2015, pp. 2,534-2,540.
- [7] L. Yang et al., “Diffusion models : A comprehensive survey of methods and applications,” *ACM Computing Surveys*, vol. 56, no. 4, pp. 1-39, November 2023.
- [8] H. Cao et al., “A survey on generative diffusion models,” *IEEE Transactions on Knowledge and Data Engineering, ISCI 2011*, Kuala Lumpur, Malaysia, March 20-23, 2011, pp. 1,532-1,540.
- [9] R. Rombach et al., “High-resolution image synthesis with latent diffusion models,” *Conference on Computer Vision and Pattern Recognition, CVPR 2022*, New Orleans, USA, June 19-24, 2022, pp. 10,684-10,695.

- [10] J. Ho et al., “Denoising diffusion probabilistic models,” *International Computer Science and Engineering Conference, ICSEC 2020*, Jawa Barat, Indonesia, December 6-12, 2020, pp. 6,840-6,851.
- [11] J. Nilsson and T. Akenine-Möller, “*Understanding SSIM*,” [Online]. Available : https://research.nvidia.com/publication/2020-07_Understanding-SSIM. [Accessed : September 29, 2025].
- [12] A. Hore and D. Ziou, “Image quality metrics : PSNR vs. SSIM,” *20th International Conference on Pattern Recognition, ICPR 2010*, Istanbul, Turkey, August 23-26, 2010, pp. 2,366-2,369.
- [13] A. Radford et al., “Learning transferable visual models from natural language supervision,” *International Conference on Machine Learning, ICML 2021*, Selangor, Malaysia, July 18-24, 2021, pp. 1,334-1,337.
- [14] R. Zhang et al., “The unreasonable effectiveness of deep features as a perceptual metric,” *Conference on Computer Vision and Pattern Recognition, CVPR 2018*, Salt Lake City, UT, USA, June 18-22, 2018, pp. 1,567-1,573.
- [15] P. Zhu et al., “Watermark-embedded adversarial examples for copyright protection against diffusion models,” *Conference on Computer Vision and Pattern Recognition, CVPR 2024*, Seattle, WA, USA, June 17-21, 2024, pp. 420-429.
- [16] S. Shan et al., “Nightshade : Prompt-specific poisoning attacks on text-to-image generative models,” *IEEE Symposium on Security and Privacy, ISCI 2024*, San Francisco, CA, USA, May 20-23, 2024, pp. 2,340-2,357.
- [17] S. Shan et al., “Anti-dreamBooth : Protecting users from personalized text-to-image synthesis,” *International Conference on Computer Vision, ICCV 2023*, Paris, France, October 1-6, 2023, pp. 2,257-2,268.
- [18] R. Chen et al., “EditShield : Protecting unauthorized image editing by instruction-guided diffusion models,” *European Conference on Computer Vision, ECCV 2024*, Milan, Italy, September October 4-7, 2024, pp. 126-142.
- [19] C. Liang et al., “Adversarial example does good : Preventing painting imitation from diffusion models via adversarial examples,” *International Conference on Machine Learning, ICML 2023*, Honolulu, HI, USA, July 23-29, 2023, pp. 123-140.

- [20] Y. Cui et al., “Diffusion shield : A watermark for data copyright protection against generative diffusion models,” *TEM Journal*, vol. 26, no. 2, pp. 60-75, January 2025.
- [21] H. Salman et al., “Raising the cost of malicious AI-powered image editing,” *International Conference on Machine Learning, ICML 2023, Honolulu, HI, USA, July 23-29, 2023*, pp. 534-540.



ภาคผนวก

TNI

ภาคผนวก ก.

เอกสารประกอบการนำเสนอการสอบป้องกันวิทยานิพนธ์

TNI

การออกแบบและพัฒนาวิธีการป้องกันภาพที่มีลิขสิทธิ์ถูกใช้ใน

Generative AI

Design and Development of Methods for Protecting Copyrighted
Images from Unauthorized Use in Generative AI Systems

บทที่ 1

บทนำ

TNI

Chapter 1

ความเป็นมาและความสำคัญของปัญหา

- Generative AI คือปัญญาประดิษฐ์เชิงสร้างสรรค์ที่สามารถสร้างข้อความ ภาพ วิดีโอ เสียง หรือโค้ดโปรแกรม
- เป็นเทคโนโลยีที่ได้รับความนิยมทั่วโลก โดยมีบริษัทเทคโนโลยีักษ์ใหญ่ เช่น Microsoft, Google, Meta, Baidu และ Alibaba ลงทุนและพัฒนาเพื่อนำไปใช้ในผลิตภัณฑ์ของตน

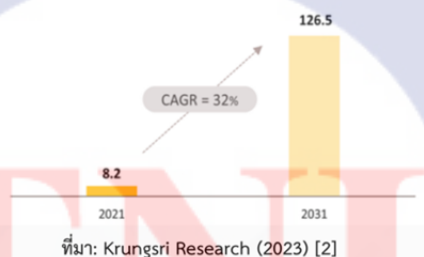


3

Chapter 1

ความเป็นมาและความสำคัญของปัญหา - 2

คาดการณ์ว่ามูลค่าตลาดของเทคโนโลยีนี้จะเติบโตอย่างก้าวกระโดด จาก 8.2 พันล้านดอลลาร์สหรัฐในปี 2021 คาดว่าจะเพิ่มเป็น 126.5 พันล้านดอลลาร์สหรัฐในปี 2031



4

Chapter 1

ความเป็นมาและความสำคัญของปัญหา - 3

Getty Images ฟ้องร้อง Stability AI (2023)

Getty Images ได้ยื่นฟ้องผู้พัฒนาโปรแกรมสร้างภาพด้วย AI ชื่อ Stable Diffusion โดยกล่าวหาว่า บริษัทนี้ได้ทำการดัดแปลงหลายล้านภาพที่มีลิขสิทธิ์จากฐานข้อมูลของ Getty ไปใช้โดยไม่ได้รับอนุญาต

Andersen v. Stability AI (2023)

เป็นการฟ้องร้องโดยกลุ่มศิลปิน นำโดย Sarah Andersen ต่อบริษัท Stability AI, Midjourney, และ DeviantArt ซึ่งเกี่ยวกับการกล่าวหาว่าบริษัทเหล่านี้ได้นำผลงานศิลปะที่มีลิขสิทธิ์ไปใช้โดยไม่ได้รับอนุญาตในการฝึกโมเดล AI แบบสร้างสรรค์

5

Chapter 1

ความเป็นมาและความสำคัญของปัญหา - 4

- Generative AI กำลังสร้างประเด็นถกเถียงสำคัญในด้าน ลิขสิทธิ์ (Copyright) และ จริยธรรม (Ethics) โดยเฉพาะกรณีที่มีการนำข้อมูลหรือภาพที่มีลิขสิทธิ์ไปใช้ในการฝึก โมเดล AI หรือการนำไปใช้โดยไม่ได้รับอนุญาต
- งานวิจัยนี้จึงมีเป้าหมายเพื่อออกแบบและพัฒนาวิธีการป้องกันภาพที่มีลิขสิทธิ์ถูกใช้ใน Generative AI โดยไม่ได้รับการอนุญาต

6

Chapter 1

วัตถุประสงค์งานวิจัย

- เพื่อพัฒนาอัลกอริทึมการป้องกันภาพที่มีลิขสิทธิ์สำหรับโมเดล Stable Diffusion 1.5 ภายใต้กระบวนการ Image-to-Image (img2img)
- เพื่อประเมินความสามารถของวิธีการในการรักษาความใกล้เคียงกับภาพต้นฉบับ ขณะเดียวกันทำให้ผลลัพธ์จากโมเดล Stable Diffusion 1.5 ภายใต้กระบวนการ Image-to-Image (img2img) เกิดความเบี่ยงเบนจากผลลัพธ์ปกติ

7

Chapter 1

ขอบเขตของการวิจัย

- 1.การวิจัยนี้มุ่งเน้นการหาวิธีป้องกันการถูกนำภาพที่มีลิขสิทธิ์ ไปใช้การสร้างภาพใหม่หรือการดัดแปลงภาพนั้นด้วย Generative AI Model โดยเฉพาะ Stable Diffusion Model
- 2.ประเภทของภาพที่ต้องการป้องกัน คือ สัตว์ สิ่งของ ยานพาหนะ ฯลฯ
- 3.ข้อมูลที่ใช้ในการทดลองเป็นชุดข้อมูลสาธารณะ เช่น ImageNet COCO Dataset และอื่นๆ

8

Chapter 1

ขอบเขตของการวิจัย

4.การวิจัยนี้จำกัดขอบเขตเฉพาะ Diffusion Model SD 1.5

5.เมื่อนำภาพที่ถูกป้องกันไปใช้กับ Diffusion Model SD 1.5 จะทำให้เกิด ผลลัพธ์ที่
มีการเบี่ยงเบนหรือผิดเพี้ยนไปจากผลลัพธ์ปกติ

9

Chapter 1

ประโยชน์ที่คาดว่าจะได้รับ

สามารถป้องกันการถูกนำภาพที่มีลิขสิทธิ์ไปใช้ใน Generative AI โดยที่ไม่ได้รับ
การยินยอมจากเจ้าของผลงาน

10

บทที่ 2

หลักการทฤษฎีและงานวิจัยที่เกี่ยวข้อง

11

Chapter 2

หลักการทฤษฎีและงานวิจัยที่เกี่ยวข้อง

- 1 ทฤษฎีที่เกี่ยวข้องกับปัญญาประดิษฐ์เชิงสร้างสรรค์ (Generative AI)
- 2 ตัวอย่างเชิงปฏิบัติ (Adversarial Examples)
- 3 แบบจำลองการแพร่กระจาย (Diffusion Model)

12

Chapter 2

หลักการทฤษฎีและงานวิจัยที่เกี่ยวข้อง

- 4 Stable Diffusion
- 5 ตัวชี้วัดการประเมินผล (Evaluation Metrics)
- 6 งานวิจัยที่เกี่ยวข้อง

13

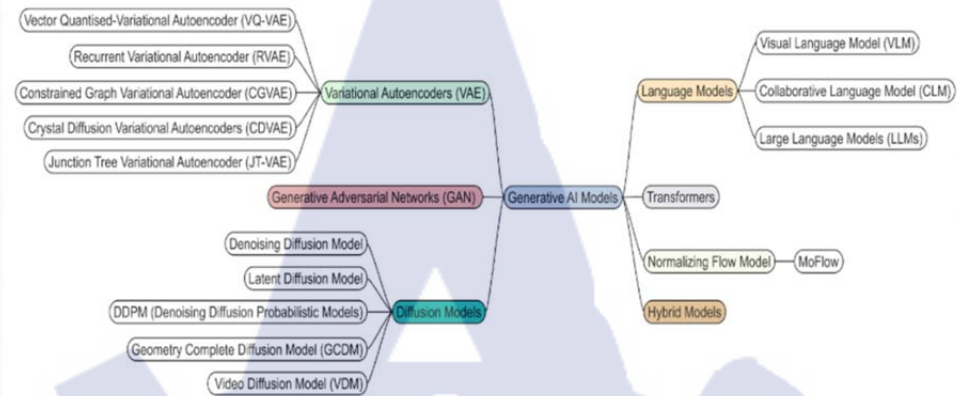
Chapter 2

ทฤษฎีที่เกี่ยวข้องกับปัญญาประดิษฐ์เชิงสร้างสรรค์ (Generative AI) - 1

- ระบบปัญญาประดิษฐ์ที่สามารถสร้างข้อมูลใหม่ได้ เช่น ข้อความ ภาพ เสียง วิดีโอ
- ใช้หลักการของ โมเดลการสร้างข้อมูล (Generative Models) ที่เรียนรู้ “รูปแบบ” และ “โครงสร้าง” ของข้อมูลจากชุดฝึก
- ตัวอย่างการใช้งาน เช่น
 - สร้างภาพจากคำบรรยาย (Text-to-Image)
 - ดัดแปลงภาพ (Image-to-Image)
 - สร้างข้อความหรือเนื้อหาใหม่ (Text Generation)

14

Chapter 2

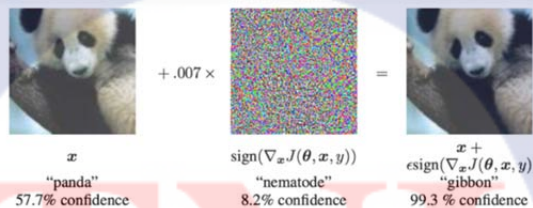


15

Chapter 2

ตัวอย่างเชิงปฏิบัติ (Adversarial Examples)

- อินพุตที่ถูกปรับแต่งด้วยสัญญาณรบกวนขนาดเล็กที่มองไม่เห็นด้วยตาเปล่า แต่ทำให้โมเดล deep learning ทำนายผลผิดพลาดได้



ภาพแพนด้าที่ถูกรบกวน (perturbation) เล็กน้อย แต่ GoogLeNet ทำนายผลผิดอย่างชัดเจน

16

Chapter 2

แบบจำลองการแพร่กระจาย (Diffusion Model) - 1

- Diffusion Models (DMs) คือแบบจำลองการสร้างข้อมูลที่อาศัยแนวคิด “การทำลายและกู้คืนข้อมูล”
- ภาพจะโดนเติมสัญญาณรบกวน (noise) เข้าไปที่ละน้อย จนข้อมูลกลายเป็น noise (forward process)
- โมเดลจะถูกฝึกให้เรียนรู้กระบวนการย้อนกลับเพื่อค่อย ๆ ลบ noise ออก จนสามารถสร้างข้อมูลใหม่ขึ้นมา (reverse process)

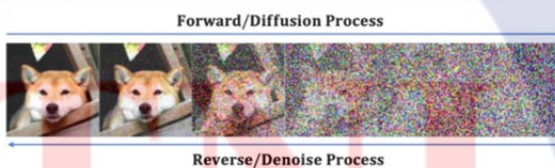
17

Chapter 2

แบบจำลองการแพร่กระจาย (Diffusion Model) - 2

การทำลายข้อมูล (Forward Process)

- โมเดลจะกำหนดวิธีการเติม noise ให้กับข้อมูลจริง
- ข้อมูลต้นฉบับ จะถูกแปลงเป็น $x_1, x_2, \dots, x_T$ ตามลำดับเวลา
- จุดประสงค์ของขั้นนี้คือการสร้าง “เส้นทาง” ที่ชัดเจนจากข้อมูลจริงไปยัง noise



18

Chapter 2

แบบจำลองการแพร่กระจาย (Diffusion Model) - 3

การกู้คืนข้อมูล (Reverse Process)

- โมเดลจะทำนายว่า noise ที่เติมเข้าไปคืออะไร แล้วค่อย ๆ ลบมันออกไปจนผลลัพธ์สุดท้ายคือได้ข้อมูลใหม่ ที่มีการแจกแจงใกล้เคียงกับข้อมูลจริง
- ถ้ามี prompt มันจะส่งเข้าไปในขั้นตอนการทำนายทุกครั้งเพื่อบอกว่าภาพที่ต้องการมีลักษณะแบบใด

19

Chapter 2

แบบจำลองการแพร่กระจาย (Diffusion Model) - 4

แบบจำลอง Diffusion ถูกนำไปใช้ในเครื่องมือสร้างสรรค์เชิงพาณิชย์หลายประเภท เช่น

- DALL-E 2 ของ OpenAI
- Stable Diffusion (SD) ของ Stability AI
- Imagen ของ Google

ซึ่งทั้งหมดนี้อาศัยกระบวนการสร้างภาพโดยมี Diffusion Model เป็นกลไกหลักในการสร้างภาพที่มีความสมจริงสูง

20

Chapter 2

Stable Diffusion - 1

- Stable Diffusion เป็นโมเดลสร้างภาพแบบที่พัฒนามาบนพื้นฐานของ Latent Diffusion Model (LDM)
- ไม่ทำ diffusion บนภาพโดยตรงใน pixel space แต่ทำงานใน latent space ที่มีมิติเล็กกว่า ซึ่งช่วยลดภาระการคำนวณอย่างมาก ใช้หน่วยความจำต่ำกว่า และสามารถสร้างภาพความละเอียดสูงได้อย่างมีประสิทธิภาพ

21

Chapter 2

Stable Diffusion - 2

Stable Diffusion ประกอบด้วย 3 ส่วนสำคัญ

- VAE ทำหน้าที่เข้ารหัสภาพจาก pixel space ไปยัง latent space และถอดรหัสกลับเป็นภาพจริง ทำให้ diffusion สามารถดำเนินการใน latent space ได้
- UNet เป็นเครือข่ายหลักที่รับ latent ที่มี noise และค่อย ๆ ทำนายและลบสัญญาณรบกวนในกระบวนการ reverse diffusion
- Cross-Attention เชื่อมโยงข้อความกับโครงสร้างภาพ โดยส่ง text embedding เข้าไปกับการทำงานของ UNet เพื่อให้ภาพสอดคล้องกับ prompt

22

Chapter 2

Stable Diffusion - 3

- ใน Stable Diffusion ค่าเริ่มต้นของ latent noise ถูกกำหนดโดย **ตัวกำเนิดตัวเลขสุ่ม (random number generator)** ซึ่งควบคุมผ่านค่าที่เรียกว่า “seed”
- หากใช้ seed เดียวกันร่วมกับพารามิเตอร์อื่นที่เหมือนกัน โมเดลจะสร้างภาพผลลัพธ์เดิมทุกครั้ง แต่หากเปลี่ยน seed จุดเริ่มต้นของกระบวนการ diffusion จะเปลี่ยนไป ส่งผลให้ได้ภาพที่แตกต่างกัน

$$z_T \sim \mathcal{N}(0, I)$$

23

Chapter 2

ตัวชี้วัดการประเมินผล (Evaluation Metrics) - 1

SSIM (Structural Similarity Index Measure)

เป็นตัวชี้วัดที่ใช้ประเมินว่า “ภาพที่ได้มีความคล้ายกับภาพต้นฉบับมากน้อยเพียงใด” โดยไม่พิจารณาเพียงความแตกต่างของ pixel แต่คำนึงถึง structure, luminance, และ contrast ของภาพรวมด้วย

$$SSIM(x, y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)}$$

24

Chapter 2

ตัวชี้วัดการประเมินผล (Evaluation Metrics) - 2

PSNR (Peak Signal-to-Noise Ratio)

เป็นตัวชี้วัดที่ใช้ประเมินว่า “ภาพที่ได้แตกต่างจากภาพต้นฉบับมากน้อยเพียงใด” โดยคำนวณจากค่า ความคลาดเคลื่อนเฉลี่ย (MSE) ซึ่งวัดจาก ผลต่างของค่าพิกเซล ระหว่างภาพทั้งสอง

$$PSNR(f, g) = 10 \log_{10} \left(\frac{MAX^2}{MSE(f, g)} \right)$$

25

Chapter 2

ตัวชี้วัดการประเมินผล (Evaluation Metrics) - 3

CLIP Similarity

เป็นตัวชี้วัดที่ใช้ดูว่า “ภาพสองภาพมีความหมายใกล้เคียงกันแค่ไหน” หากค่า similarity สูง แสดงว่าภาพทั้งสองมีความหมายหรือเนื้อหาคล้ายกันในเชิง semantics

LPIPS (Learned Perceptual Image Patch Similarity)

เป็นตัวชี้วัดที่ใช้ประเมินความใกล้เคียงของภาพสองภาพในเชิงการรับรู้ของมนุษย์ โดยใช้คุณลักษณะ (feature) ที่สกัดจากโครงข่ายประสาทเชิงลึกที่ผ่านการฝึกมาแล้วในการคำนวณความแตกต่างของภาพ

26

Chapter 2

งานวิจัยที่เกี่ยวข้อง - 1

Nightshade

- Nightshade คือเทคนิคการโจมตีแบบ data-poisoning ที่ใช้ประโยชน์จากความเบาบางของการแทนความคิดเชิงคอนเซ็ปต์ (concept sparsity) ที่เชื่อมโยงคอนเซ็ปต์เป้าหมายกับคอนเซ็ปต์อื่น เช่น ผูกคำว่า “สุนัข” ให้เกิดภาพ “แมว”
- เมื่อโมเดลถูกฝึกด้วยข้อมูลนี้จะเรียนรู้แบบผิดเพี้ยน ทำให้เมื่อเรียกใช้งานด้วยคำสั่งที่เกี่ยวกับคอนเซ็ปต์เป้าหมาย โมเดลมักจะสร้างภาพของคอนเซ็ปต์ที่ถูกวางยาแทน

27

Chapter 2

งานวิจัยที่เกี่ยวข้อง - 2

EditShield

- ป้องกันการแก้ไขภาพโดย AI ที่ใช้คำสั่งแบบข้อความ (Instruction-guided)
- ใช้การเพิ่ม สัญญาณรบกวน (Perturbation) เพื่อรบกวนการเข้ารหัสใน Latent Space
- ทำให้โมเดล Diffusion ดีความภาพต้นฉบับผิด ตั้งแต่ขั้นตอนแรกของการสร้าง
- เมื่อมีการสั่งแก้ไข ผลลัพธ์จะ ผิดเพี้ยนหรือแก้ไขไม่สำเร็จ

28

Chapter 2

งานวิจัยที่เกี่ยวข้อง - 3

PhotoGuard

- เป็นเทคนิคการป้องกันการแก้ไขภาพ ผ่านการเพิ่มสัญญาณรบกวนขนาดเล็กที่มองไม่เห็น (imperceptible perturbation)
- ใช้แนวคิดการโจมตีเชิงปฏิบัติ (Adversarial Attack) ด้วยการทำให้ **Latent** ของภาพไม่มีความหมาย เพื่อรบกวนการเข้ารหัสและกระบวนการสร้างภาพในโมเดล
- ทำให้โมเดล Diffusion ไม่สามารถแก้ไขภาพ (img2img / inpainting) ได้อย่างถูกต้องหรือสมจริง ขณะที่ภาพต้นฉบับยังคงมีลักษณะปกติ

29

Chapter 2

งานวิจัยที่เกี่ยวข้อง - 4

Watermark-embedded Adversarial Examples for Copyright Protection

- สร้าง “ตัวอย่างปฏิบัติ (Adversarial Example)” ที่มีการฝังลายน้ำ (Watermark) ลงในสัญญาณรบกวน
- การรบกวนเชิงปฏิบัติถูกออกแบบเพื่อขยับการแทนเชิงแฝง (**latent representation**) ของภาพ ให้เข้าใกล้การแทนเชิงแฝงของลายน้ำ
- บังคับให้ Diffusion Model สร้างภาพที่มีลายน้ำของผู้สร้างปรากฏชัดเจน เพื่อแสดงความเป็นเจ้าของผลงาน

30

นิยามคำศัพท์

- Latent คือ ตัวแทนความหมายข้อมูลในโลกของโมเดล หรือก็คือ “สิ่งที่โมเดลเข้าใจ” ว่าภาพนั้นมีอะไรอยู่บ้าง เช่น สี รูปร่าง วัตถุ บริบท หรืออารมณ์ของภาพ
- Seed ค่าตัวเลขที่ใช้กำหนด จุดเริ่มต้นของกระบวนการสุ่ม ในการสร้างภาพของโมเดล ทำให้สามารถควบคุมความสุ่มและสร้างผลลัพธ์ซ้ำเดิมได้
- Perturbation การเพิ่ม การรบกวนขนาดเล็ก ลงใน ข้อมูลหรือ latent ของภาพ เพื่อให้เกิดการเปลี่ยนแปลงของผลลัพธ์จากโมเดล

31

บทที่ 3

ระเบียบและวิธีดำเนินการวิจัย

TNI

32

Chapter 3

ขั้นตอนการดำเนินงาน

- 1 วิเคราะห์ข้อมูลและปัญหาของงานวิจัย
- 2 การศึกษาข้อมูลและเทคโนโลยีที่ต้องนำมาใช้
- 3 การออกแบบการทดลอง
- 4 การดำเนินการทดลอง
- 5 การประเมินผลการวิจัย

33

Chapter 3

วิเคราะห์ข้อมูลและปัญหาของงานวิจัย

- ปัจจุบันมีการใช้ ปัญญาประดิษฐ์เชิงสร้างสรรค์ (Generative AI) โดยเฉพาะ Diffusion Model ในการสร้างและดัดแปลงภาพอย่างแพร่หลาย
- เกิดปัญหาการนำภาพที่มีลิขสิทธิ์ไปใช้โดยไม่ได้รับอนุญาต
- จากการศึกษาปัญหาและแนวทางป้องกันดังกล่าว ผู้วิจัยจึงเลือกทำหัวข้อ “การออกแบบและพัฒนาวิธีการป้องกันภาพที่มีลิขสิทธิ์ถูกใช้ใน Generative AI”

34

Chapter 3

การศึกษาข้อมูลและเทคโนโลยีที่ต้องนำมาใช้

ผู้วิจัยศึกษาข้อมูลและเทคโนโลยีที่เกี่ยวข้องกับการทำงานวิจัยโดยแบ่งออกเป็น 4 ส่วนหลักได้แก่

- แบบจำลองการแพร่กระจาย (Diffusion Model)
- Stable Diffusion
- การรบกวนเชิงแฝง (Perturbation)
- ตัวชี้วัดการประเมิน Evaluation Metrics

35

Chapter 3

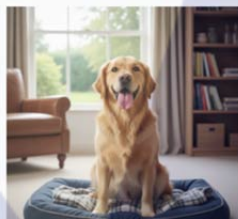
การออกแบบการทดลอง

TNI

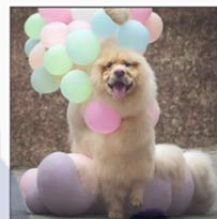
36

แนวทางการวิจัย

- การทำให้ Stable Diffusion เข้าใจภาพ ผิดเพี้ยน ไปจากเดิม โดยการขยับ latent ของภาพต้นฉบับ ให้ไปหา latent ของภาพเป้าหมายที่ผิดเพี้ยน



Original



Target

37

Chapter 3

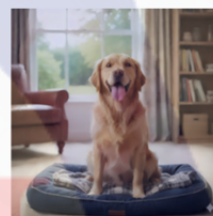
ขยับ latent เข้าหา
ภาพเป้าหมาย



ภาพต้นฉบับ



ภาพเป้าหมาย

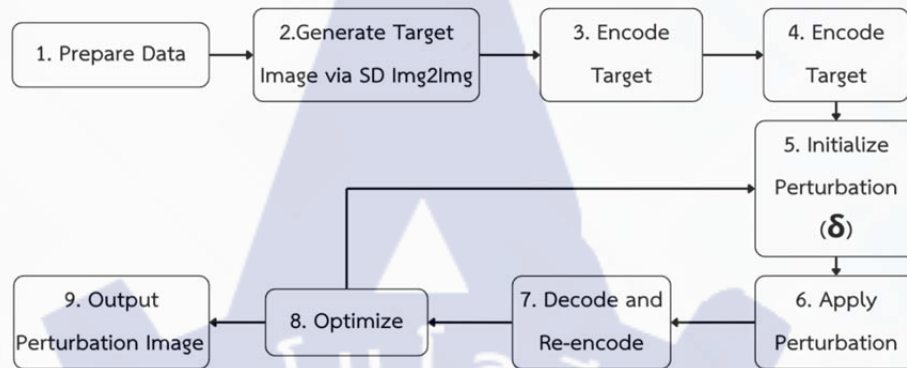


ภาพที่ผ่านการป้องกัน

38

Chapter 3

ภาพรวมขั้นตอนการออกแบบภาพที่ถูกรบกวน (Perturbation)



39

Chapter 3

การออกแบบภาพที่ถูกรบกวน - 1



- เตรียมภาพต้นฉบับ (Original)
- นำภาพต้นฉบับ เข้าสู่ Stable Diffusion 1.5 เพื่อสร้าง ภาพเป้าหมาย (target) โดยกำหนด prompt ที่ทำให้ภาพเป้าหมายมีความผิดเพี้ยนไปจากเดิม

40

Chapter 3

การออกแบบภาพที่ถูกรบกวน - 2



41

Chapter 3

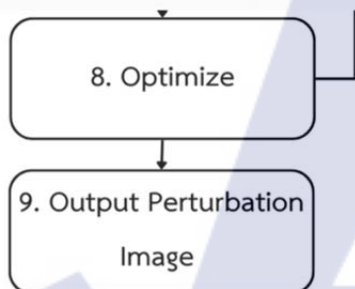
การออกแบบภาพที่ถูกรบกวน - 3



42

Chapter 3

การออกแบบภาพที่ถูกรบกวน - 3



- ทำการปรับค่า δ ด้วยวิธีการเพิ่มประสิทธิภาพเชิงเกรเดียนต์อย่างต่อเนื่อง เพื่อให้ค่าที่ได้มีความเหมาะสม
- เมื่อกระบวนการปรับค่าเสร็จสมบูรณ์ จะได้ภาพที่ผ่านการฝัง perturbation ซึ่งมีลักษณะใกล้เคียงกับภาพต้นฉบับ แต่สามารถทำให้ผลลัพธ์ที่สร้างจากโมเดลเกิดความเบี่ยงเบนได้

43

Chapter 3

การดำเนินการทดลอง

TNI

44

Chapter 3

การดำเนินการทดลอง - 1

- ผู้วิจัยใช้ชุดข้อมูลภาพต้นฉบับจำนวน 20 ภาพ แบ่งเป็น 4 หมวดหมู่ ได้แก่ สัตว์ สิ่งของ สถานที่ และบุคคล เพื่อทดสอบวิธีการกับภาพที่มีลักษณะแตกต่างกัน
- ภาพเป้าหมาย (Target Image) ถูกสร้างด้วยกระบวนการ image-to-image ของโมเดล Stable Diffusion 1.5 โดยใช้ prompt โครงสร้างเดียวกันทุกภาพ โดยเปลี่ยนเฉพาะชนิดของวัตถุ เพื่อให้ภาพที่ได้มีลักษณะผิดปกติในทิศทางเดียวกัน เช่น โทนีสี pastel pink รูปร่างพองกลม ลักษณะคล้ายลูกโป่ง และสัดส่วนที่ผิดเพี้ยน

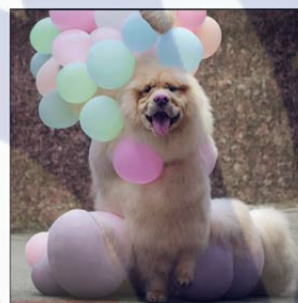
45

Chapter 3



ภาพต้นฉบับ

Stable
Diffusion 1.5
→



ภาพเป้าหมาย

46

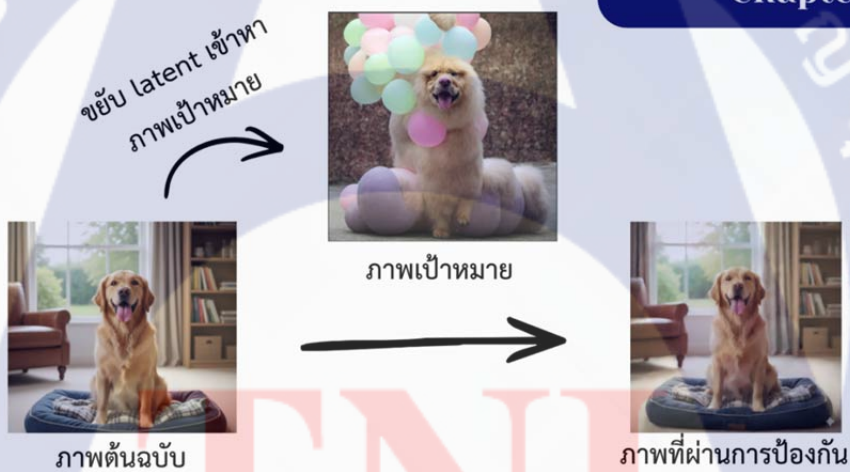
Chapter 3

การดำเนินการทดลอง - 2

- คำนวณค่าการรบกวนเชิงแฝง (δ) สำหรับภาพแต่ละคู่ (Original-Target) ด้วยกระบวนการ Gradient-based Optimization ภายใต้ฟังก์ชันความสูญเสีย เพื่อหา perturbation ที่เหมาะสม
- นำค่า δ ที่ผ่านการ Optimization ร่วมกับ latent ของภาพต้นฉบับ แล้ว decode กลับเป็นภาพใหม่ (Perturbed Image) ซึ่งมีลักษณะใกล้เคียงกับภาพต้นฉบับ แต่แฝงการรบกวนไว้ภายใน

47

Chapter 3



48

Chapter 3

การดำเนินการทดลอง - 3

- นำภาพต้นฉบับ และ ภาพที่ผ่านการป้องกัน เข้าสู่โมเดล Stable Diffusion 1.5 ในโหมด img2img โดยใช้ พารามิเตอร์และ prompt เดียวกัน
- เนื่องจากกระบวนการ diffusion มีลักษณะเชิงสุ่มตามค่า random seed ผู้วิจัยจึงกำหนดให้ใช้ seed = 3333 เพื่อควบคุมเงื่อนไขการสร้างภาพให้คงที่

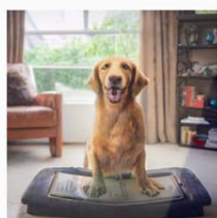
49



50

Chapter 3

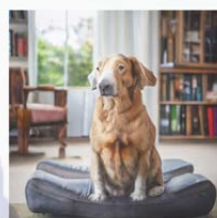
การดำเนินการทดลอง - 4



seed 100



seed 500



seed 3333

51

Chapter 3

การประเมินผลการวิจัย

1. การประเมินคุณภาพของภาพต้นฉบับและภาพที่ผ่านการป้องกัน
2. การประเมินผลกระทบของภาพที่ผ่านการป้องกันต่อกระบวนการสร้างภาพของโมเดล

SSIM

CLIP Similarity

PSNR

LPIPS

52

บทที่ 4 ผลการวิจัย

53

Chapter 4

การประเมินผล

ตัวชี้วัดที่ใช้ ได้แก่ SSIM, PSNR, CLIP Similarity และ LPIPS โดยแบ่งการประเมินเป็นสองส่วน

- การประเมินคุณภาพของภาพต้นฉบับและภาพที่ผ่านการป้องกัน
- การประเมินผลกระทบของภาพที่ผ่านการป้องกันต่อกระบวนการสร้างภาพของโมเดล

54

Chapter 4

การประเมินคุณภาพของภาพต้นฉบับ
และภาพที่ผ่านการป้องกัน

55

Chapter 4

Image	SSIM	PSNR	Clip	LPIPS
Dog	0.8608	27.8	0.9624	0.1579
Cat	0.8874	27.95	0.9599	0.1485
Rabbit	0.7203	25.18	0.9882	0.206
Parrot	0.8128	25.22	0.9669	0.1782
Butterfly	0.8434	27.43	0.9916	0.1519
Glass	0.8429	27.38	0.9675	0.1657
Lamp	0.6717	26.59	0.959	0.3338
Phone	0.8055	27.52	0.9694	0.2369
Pillow	0.7888	27.52	0.9356	0.2468
Balloon	0.8454	27.49	0.9822	0.0978

56

Chapter 4

Image	SSIM	PSNR	Clip	LPIPS
Home	0.8063	25.87	0.9858	0.1456
Garden	0.4228	19.95	0.8984	0.4229
Beach	0.7921	28.42	0.9657	0.1986
Parking	0.6276	23.8	0.9523	0.2729
Street	0.735	26.23	0.9351	0.2577
Person1	0.8548	27.49	0.9455	0.1502
Person2	0.8467	27.99	0.9036	0.1294
Person3	0.834	27.36	0.9734	0.1329
Person4	0.6729	27.4	0.9153	0.304
Person5	0.8239	27.27	0.9474	0.1241
Average	0.7748	26.59	0.9553	0.2031

57

Chapter 4

- จากค่าเฉลี่ยของตัวชี้วัดพบว่า SSIM มีค่าเฉลี่ยเท่ากับ 0.7748 และ PSNR เท่ากับ 26.59 dB แสดงให้เห็นว่าภาพที่ผ่านการป้องกันยังคงรักษาโครงสร้างและคุณภาพเชิงพิกเซลของภาพต้นฉบับไว้ได้ในระดับที่ยอมรับได้
- ค่า CLIP Similarity มีค่าเฉลี่ยสูงถึง 0.9553 สะท้อนว่าภาพยังคงถูกตีความว่าเป็นวัตถุประเภทเดิมในเชิงความหมาย ส่วนค่า LPIPS มีค่าเฉลี่ยอยู่ที่ 0.2031 ซึ่งอยู่ในระดับต่ำถึงปานกลาง บ่งชี้ว่าความแตกต่างเชิงการรับรู้ของมนุษย์อยู่ในระดับที่ไม่เด่นชัด

58

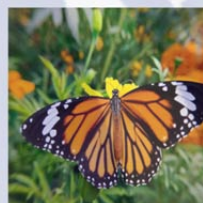
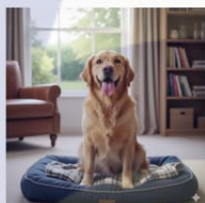
Chapter 4

ภาพต้นฉบับและภาพที่ผ่านการป้องกันในกลุ่ม
ตัวอย่างประเภทสัตว์

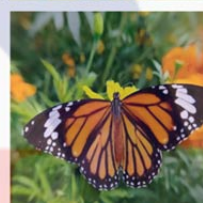
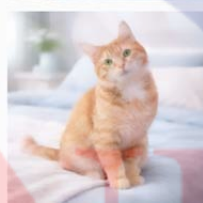
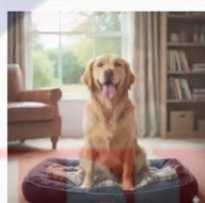
59

Chapter 4

ภาพต้นฉบับ



ภาพที่ผ่านการป้องกัน



Dog

Cat

Butterfly

60

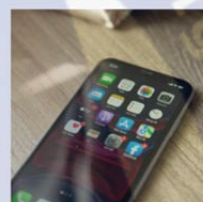
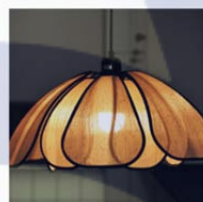
Chapter 4

ภาพต้นฉบับและภาพที่ผ่านการป้องกันในกลุ่ม
ตัวอย่างประเภทสิ่งของ

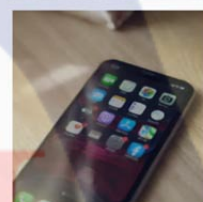
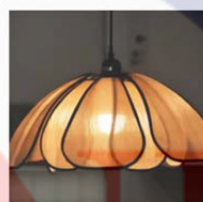
61

Chapter 4

ภาพต้นฉบับ



ภาพที่ผ่านการป้องกัน



Glass

Lamp

Phone

62

Chapter 4

ภาพต้นฉบับและภาพที่ผ่านการป้องกันในกลุ่ม
ตัวอย่างประเภทสถานที่

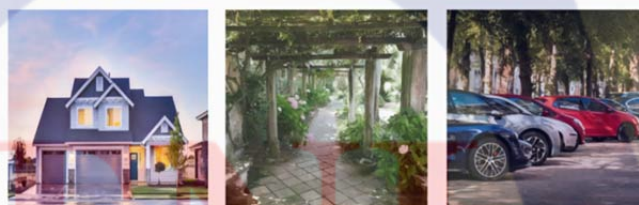
63

Chapter 4

ภาพต้นฉบับ



ภาพที่ผ่านการป้องกัน



Home

Garden

Parking

64

Chapter 4

ภาพต้นฉบับและภาพที่ผ่านการป้องกันในกลุ่ม
ตัวอย่างประเภทยบุคคล

65

Chapter 4

ภาพต้นฉบับ



ภาพที่ผ่านการป้องกัน



Person1

Person2

Person3

66

Chapter 4

การประเมินผลกระทบบของภาพที่ผ่านการป้องกัน
ต่อกระบวนการสร้างภาพของโมเดล

67

Chapter 4

Image	SSIM	PSNR	Clip	LPIPS
Dog	0.5056	15.91	0.8227	0.5365
Cat	0.6859	16.61	0.8799	0.4741
Rabbit	0.3487	15.97	0.9293	0.4858
Parrot	0.6306	16.03	0.9163	0.3688
Butterfly	0.4647	14.5	0.8878	0.577
Glass	0.6307	16.04	0.8902	0.5012
Lamp	0.4893	16.94	0.8007	0.6466
Phone	0.623	18.04	0.9248	0.5045
Pillow	0.6319	19.36	0.891	0.5502
Balloon	0.7333	19.8	0.9118	0.2717

68

Chapter 4

Image	SSIM	PSNR	Clip	LPIPS
Home	0.593	15.97	0.923	0.3862
Garden	0.0936	11.83	0.7101	0.6408
Beach	0.7016	22.5	0.8314	0.3788
Parking	0.2503	13.84	0.8435	0.5665
Street	0.4546	16.63	0.7281	0.5673
Person1	0.6497	18.09	0.6566	0.4493
Person2	0.6836	19.13	0.4389	0.4558
Person3	0.6714	18.26	0.7627	0.3183
Person4	0.5035	19.88	0.8532	0.4718
Person5	0.6531	17.52	0.7045	0.3684
Average	0.5499	17.14	0.8153	0.476

69

Chapter 4

- จากค่าเฉลี่ยของตัวชี้วัดพบว่า SSIM = 0.5390 และ PSNR = 17.37 dB แสดงให้เห็นว่าภาพที่สร้างจากภาพต้นฉบับ (Ori_img2img) และภาพที่สร้างจากภาพที่ผ่านการป้องกัน (Pert_img2img) ภายใต้กระบวนการ image-to-image มีความแตกต่างเชิงโครงสร้างและเชิงพิกเซลในระดับค่อนข้างสูง
- ในด้านความหมายและการรับรู้ ค่า CLIP Similarity = 0.8027 บ่งชี้ว่ายังคงรักษาความหมายหลักของวัตถุได้บางส่วน ขณะที่ค่า LPIPS = 0.4718 แสดงว่าภาพทั้งสองมีความแตกต่างเชิงการรับรู้ของมนุษย์อย่างชัดเจน

70

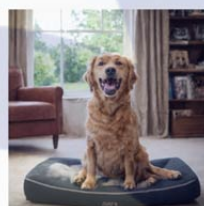
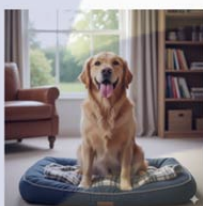
Chapter 4

ตัวอย่างผลลัพธ์ของภาพที่ผ่าน Stable Diffusion 1.5 ของภาพ
ต้นฉบับและภาพที่ผ่านการป้องกัน ในกลุ่มตัวอย่างประเภทสัตว์

71

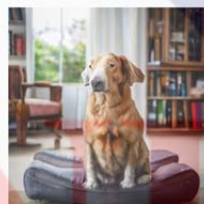
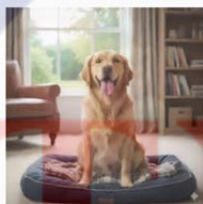
Chapter 4

ภาพต้นฉบับ



ภาพที่สร้างจากภาพต้นฉบับ

ภาพที่ผ่านการป้องกัน

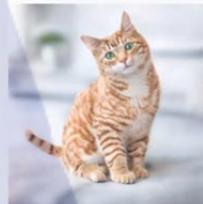
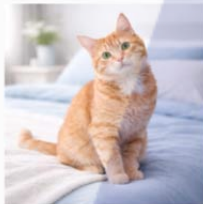


ภาพที่สร้างจากภาพที่ผ่าน
การป้องกัน

72

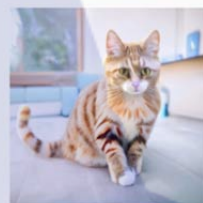
Chapter 4

ภาพต้นฉบับ



ภาพที่สร้างจากภาพต้นฉบับ

ภาพที่ผ่านการป้องกัน

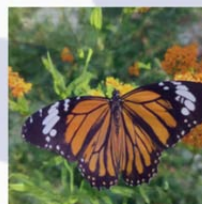
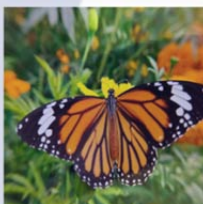


ภาพที่สร้างจากภาพที่ผ่านการป้องกัน

73

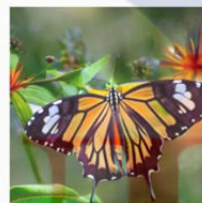
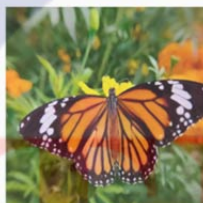
Chapter 4

ภาพต้นฉบับ



ภาพที่สร้างจากภาพต้นฉบับ

ภาพที่ผ่านการป้องกัน



ภาพที่สร้างจากภาพที่ผ่านการป้องกัน

74

Chapter 4

ตัวอย่างผลลัพธ์ของภาพที่ผ่าน Stable Diffusion 1.5 ของภาพ
ต้นฉบับและภาพที่ผ่านการป้องกัน ในกลุ่มตัวอย่างประเภทสิ่งของ

75

Chapter 4

ภาพต้นฉบับ



ภาพที่สร้างจากภาพต้นฉบับ

ภาพที่ผ่านการป้องกัน

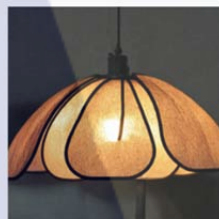
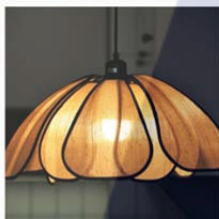


ภาพที่สร้างจากภาพที่ผ่าน
การป้องกัน

76

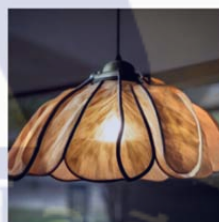
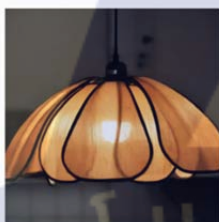
Chapter 4

ภาพต้นฉบับ



ภาพที่สร้างจากภาพต้นฉบับ

ภาพที่ผ่านการป้องกัน



ภาพที่สร้างจากภาพที่ผ่าน
การป้องกัน

77

Chapter 4

ภาพต้นฉบับ



ภาพที่สร้างจากภาพต้นฉบับ

ภาพที่ผ่านการป้องกัน



ภาพที่สร้างจากภาพที่ผ่าน
การป้องกัน

78

Chapter 4

ตัวอย่างผลลัพธ์ของภาพที่ผ่าน Stable Diffusion 1.5 ของภาพต้นฉบับและภาพที่ผ่านการป้องกัน ในกลุ่มตัวอย่างประเภทสถานที่

79

Chapter 4

ภาพต้นฉบับ



ภาพที่สร้างจากภาพต้นฉบับ



ภาพที่ผ่านการป้องกัน



ภาพที่สร้างจากภาพที่ผ่านการป้องกัน



80

Chapter 4

ภาพต้นฉบับ



ภาพที่สร้างจากภาพต้นฉบับ



ภาพที่ผ่านการป้องกัน



ภาพที่สร้างจากภาพที่ผ่านการป้องกัน



81

Chapter 4

ภาพต้นฉบับ



ภาพที่สร้างจากภาพต้นฉบับ



ภาพที่ผ่านการป้องกัน



ภาพที่สร้างจากภาพที่ผ่านการป้องกัน



82

Chapter 4

ตัวอย่างผลลัพธ์ของภาพที่ผ่าน Stable Diffusion 1.5 ของภาพต้นฉบับและภาพที่ผ่านการป้องกัน ในกลุ่มตัวอย่างประเภทบุคคล

83

Chapter 4

ภาพต้นฉบับ



ภาพที่สร้างจากภาพต้นฉบับ



ภาพที่ผ่านการป้องกัน



ภาพที่สร้างจากภาพที่ผ่านการป้องกัน



84

Chapter 4

ภาพต้นฉบับ



ภาพที่สร้างจากภาพต้นฉบับ

ภาพที่ผ่านการป้องกัน



ภาพที่สร้างจากภาพที่ผ่านการป้องกัน

85

Chapter 4

ภาพต้นฉบับ



ภาพที่สร้างจากภาพต้นฉบับ

ภาพที่ผ่านการป้องกัน



ภาพที่สร้างจากภาพที่ผ่านการป้องกัน

86

บทที่ 5

สรุปผลและข้อเสนอแนะ

87

Chapter 5

สรุปผลการวิจัย

- ภาพที่ผ่านการป้องกันยังคง รักษาโครงสร้าง คุณภาพเชิงพิกเซล และความหมายใกล้เคียงภาพต้นฉบับ ในระดับที่ยอมรับได้
- เมื่อใช้ภาพดังกล่าวในกระบวนการ image-to-image ของ Stable Diffusion 1.5 ผลลัพธ์ที่ได้มี ความแตกต่างจากภาพที่สร้างจากภาพต้นฉบับอย่างชัดเจน ทั้งเชิงโครงสร้าง เชิงความหมาย และเชิงการรับรู้
- การวิเคราะห์เชิงคุณภาพพบว่า perturbation ส่งผลต่อภาพกลุ่มสัตว์และบุคคลเด่นชัด ขณะที่ภาพประเภทวัตถุและสถานที่บางกรณียังคงรักษาโครงสร้างหลักได้

88

Chapter 5

ปัญหาและข้อจำกัด - 1

- ในบางกรณีการฝัง perturbation อาจทำให้รายละเอียดของภาพต้นฉบับลดลง หรือเกิดความผิดปกติบางส่วน สะท้อนถึงความท้าทายในการรักษาสมดุลระหว่าง คุณภาพของภาพต้นฉบับ และ ความสามารถในการรบกวนผลลัพธ์ของโมเดล
- ประสิทธิภาพของ perturbation แตกต่างกันตามประเภทของภาพ โดย ภาพกลุ่มสัตว์ และบุคคลแสดงผลกระทบชัดเจนกว่า ขณะที่ภาพประเภทสิ่งปลูกสร้างและวัตถุทั่วไปยังคงรักษาโครงสร้างหลักได้ในบางกรณี

89

Chapter 5

ปัญหาและข้อจำกัด - 2

- กระบวนการสร้าง perturbation ต้องใช้การปรับค่าแบบวนซ้ำร่วมกับการคำนวณ gradient และการ encode-decode ผ่าน VAE หลายรอบ ทำให้ ใช้เวลาในการประมวลผลค่อนข้างสูง
- การทดลองในงานวิจัยนี้จำกัดอยู่ที่ Stable Diffusion 1.5 และกระบวนการ image-to-image จึงยังไม่สามารถสรุปผลได้กับโมเดล diffusion อื่นหรือเงื่อนไขการสร้างภาพที่แตกต่างกัน

90

Chapter 5

ข้อเสนอแนะ

- ควรทดสอบวิธีการฝัง perturbation กับโมเดล diffusion รุ่นใหม่ เช่น Stable Diffusion XL เพื่อประเมินความสามารถในการถ่ายโอนผล (transferability) ข้ามโมเดล
- ควรพัฒนา perturbation ที่สามารถปรับตามประเภทของวัตถุ เพื่อเพิ่มความสม่ำเสมอและประสิทธิภาพของการป้องกันภาพ
- ควรศึกษาความทนทานของ perturbation ต่อ การแปลงภาพ เช่น การบีบอัด การปรับขนาด หรือการเปลี่ยนคุณภาพไฟล์

TNI

THAI - NICHI INSTITUTE OF TECHNOLOGY