Towards Robust Control in Visual Generation and Manipulation

by

Max Ku

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Master of Mathematics
in
Computer Science

Waterloo, Ontario, Canada, 2024

© Max Ku 2024

Author's Declaration

This thesis consists of material all of which I authored or co-authored: see Statement of Contributions included in the thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Statement of Contributions

This thesis include the materials which are presented at ICLR 2024 [51] and ACL 2024 [50], and also a preprint on ArXiv [52].

I conducted most of the experiments and wrote the first drafts of these papers. My supervisor Wenhu Chen contributed ideas, feedback and editing. Other co-authors either help to run experiments or polish the paper. I understand that my thesis may be made electronically available to the public.

Abstract

The fast development of generative models has started a new era in AI, especially in conditional image synthesis. Since the rise of diffusion models, current models can perform image generation with high fidelity and diversity. This thesis is driven towards controllable generation and manipulation in the image and video domains, guided by the three studies: ImagenHub's role in identifying the controllability of current state-of-the-art image synthesis models, VIEScore to produce explainable metrics in image synthesis tasks, and AnyV2V's role in performing precise video editing.

The first part of this study highlights the evaluation of the image domain. ImagenHub, which tackles the challenge of distinguishing current research to find the best working methods, also standardized the human-centered evaluation in image synthesis research. Complementarily, VIEScore act as a new explainable metric to mimicking human-like evaluation across all conditional image synthesis tasks with multimodal LLMs, tickling the scalability issue of ImagenHub.

The second study focuses on the video domain, which introduces AnyV2V, the first framework to treat video editing as an image editing problem. It leverages the editing power from off-the-shelf image editing models and the generalization power from image-to-video models to perform precise video editing. Such paradigm is training-free and allows video edits in a wide range of applications. Most importantly, we reported the increase in performance when plugging with stronger image-to-video models, highlighting the capacity of AnyV2V for adaptive evolution.

These studies form the basis of this thesis, driving toward robust control in visual generation and manipulation. Through a thorough analysis of ImagenHub and VIEScore, this research not only identifies the current capabilities and limitations of image synthesis models but also sets the stage for future advancements in evaluating image synthesis models. Then with AnyV2V, we align the image editing and video editing problem with image-to-video models, lays the groundwork for future advancements in making video editing more controllable and robust.

Acknowledgements

I would like to express my sincere gratitude to my supervisor, Prof. Wenhu Chen, for his constant and insightful guidance throughout my journey in pursuing my master's degree. His support has been instrumental in my growth, both academically and mentally. I also want to thank all my lab-mates and co-authors for fostering such a positive and collaborative research environment, making it a pleasure to work with them.

Additionally, I want to extend my heartfelt appreciation to my family for their unwavering support and encouragement. Their belief in me has been a constant source of motivation and strength, helping me overcome challenges and stay focused on my goals.

Table of Contents

Author's Declaration ii

Statement of Contributions iii

Abstract iv

Acknowledgements v

List of Figures ix

List of Tables xi

1 Introduction 1

2 Background 5

2.1 Image Generation and Editing 5

2.1.1 Text-Guided Image Generation 6

2.1.2 Control-guided Image Generation 6

2.1.3 Mask-guided Image Editing 6

2.1.4 Text-Guided Image Editing 6

2.1.5 Subject-driven Image Generation 7

2.1.6 Subject-driven Image Editing 7

2.2 Evaluation Metrics for Image Synthesis Tasks . . . 7
2.2.1 Algorithm-based evaluation . . . 7
2.2.2 Human-centered evaluation . . . 8
2.3 Video Generation and Editing . . . 8
2.3.1 Text-To-Video (T2V) Generation . . . 9
2.3.2 Image-To-Video (I2V) Generation . . . 9
2.3.3 Video-To-Video (V2V) Generation . . . 10

3 Identifying Effective Methods in Conditional Image Synthesis 11
3.1 Methodology Design . . . 12
3.1.1 ImagenHub Data . . . 12
3.1.2 ImagenHub Inference Library . . . 12
3.1.3 ImagenHub Evaluation . . . 16
3.2 Experiments Results . . . 19
3.3 Insights from Human Inspection . . . 20
3.4 Key Takeaway . . . 22

4 Explainable Metric for Evaluating Synthesized Image 25
4.1 Methodology Design . . . 26
4.1.1 Method of VIEScore . . . 26
4.1.2 Experiment Setup . . . 27
4.2 Experiment Results . . . 29
4.2.1 Correlation Study against different backbones . . . 29
4.2.2 Correlation Study against traditional metrics . . . 30
4.3 Insights and Challenges . . . 32
4.4 Key Takeaway . . . 33

5 Bringing Image Editing into Video Editing **34**
5.1 Methodology Design . . . 35
5.1.1 Flexible First Frame Editing . . . 35
5.1.2 Structural Guidance using DDIM Inverison . . . 35
5.1.3 Appearance Guidance via Spatial Feature Injection . . . 36
5.1.4 Motion Guidance through Temporal Feature Injection . . . 36
5.1.5 Overall Formulation . . . 37
5.1.6 Experiment Setup . . . 37
5.2 Experiment Results . . . 38
5.2.1 Evaluation across Applications . . . 39
5.2.2 Ablation Study . . . 44
5.3 Limitation and Remedies . . . 47
5.4 Key Takeaway . . . 48

6 Conclusion **50**

References **52**

APPENDICES **63**

A Supplementary material of ImagenHub **64**
A.1 Visualization Tools from ImagenHub . . . 64

B Supplementary material of VIEScore **67**
B.1 Prompt Templates . . . 67
B.2 Correlations . . . 72

C Supplementary material of AnyV2V **76**
C.1 Preliminary . . . 76

List of Figures

1.1 ImagenHub Data contains evaluation inputs. With the use of ImagenHub inference library, models generate outputs in a fair setting. The ImagenHub evaluators evaluate images according to the two carefully designed metrics. 2

1.2 The difference between VIEScore and traditional metrics. 3

1.3 AnyV2V adapt image editing solutions into video editing. 4

3.1 Objects presented in each task's prompt text. 13

3.2 While the mean score stayed stable, the reliability measured by Krippendorff's Alpha diminished slightly as the number of raters increased, due to added variance. 18

3.3 Overall score of the seven tasks in ImagenHub. 22

3.4 The SC and PQ score for each model in all categories. 23

4.1 VIEScore takes all the conditions and the synthesized image as input. Then the input is fed to the MLLM with the instruction prompts, producing desired metrics aligned with the instruction prompt. 26

4.2 The experiment study the Spearman correlation between MLLMs and human perspectives on rating images across all tasks in ImagenHub. 28

4.3 Backbone comparison of VIEScore in terms of correlation. 29

4.4 Comparison of VIEScore and traditional in terms of correlation by each task. 30

4.5 MLLM confused by the example image when rating images. 32

4.6 MLLM failed to distinguish minor differences in the images. 33

5.1 AnyV2V pipeline. 39

5.2 AnyV2V is robust in a wide range of prompt-based editing tasks while preserving the background. The results align the most with the text prompt and maintain high motion consistency. 41
5.3 With different image editing models, AnyV2V can achieve a wide range of editing tasks, including reference-based style transfer, subject-driven editing, and identity manipulation. 43
5.4 Visual comparisons after disabling temporal feature injection (T.I.), spatial feature injection (S.I.) and DDIM inverted initial noise (D.I.). 45
5.5 Hyperparameter study of Spatial Feature injection. 46
5.6 Hyperparameter study of Temporal Feature injection. 47
5.7 Performance of Inversion approach with AnyV2V on 128 frames edit. . . . 48

A.1 The webpage design of Imagen Museum (Text-Guided Image Generation page). 65
A.2 The webpage design of Imagen Museum (Text-Guided Image Editing page). 66

List of Tables

3.1 Evaluated models and their parameter size and runtime. 15

3.2 Rating guideline for computing the SC and PQ score. 17

3.3 Comparison on overall human score computation setting. 18

3.4 Ablation study to understand the impact of different evaluation settings. . 19

3.5 ImagenHub leaderboard for seven core tasks. 24

4.1 Ranking judgment from each metric method. $d_{SF}(r, r_*)$ is the Spearman's footrule and $\rho_S(r, r_*)$ is the Spearman's rho (correlation). LPIPS is not available for the first task because there are no reference images. 31

5.1 Quantitative comparisons for AnyV2V with baselines on prompt-based video editing. Alignment: prompt alignment; Overall: overall preference. **Bold**: best results; Underline: top-2. 39

5.2 Comparisons for three I2V models under AnyV2V framework on novel video editing tasks. Align: reference alignment; Overall: overall preference. **Bold**: best results; Underline: top-2. 42

5.3 Ablation study results for AnyV2V. T. Injection and S. Injection correspond to temporal and spatial feature injection mechanisms, respectively. 45

B.1 Correlations across all tasks with different backbone models. We highlight the highest correlation numbers in green. 72

B.2 Correlations to human ratings in different available methods. (Part 1) . . . 73

B.3 Correlations to human ratings in different available methods. (Part 2) . . . 74

B.4 Correlations to human ratings in different available methods. (Part 3) . . . 75

Chapter 1

Introduction

In the rapidly evolving field of digital content creation with generative models, visual content generation and manipulation have seen significant advancements [109, 84, 42, 43, 44]. This thesis focuses on the image and video domain generation and editing. The first part focuses on the evaluation of image synthesis, where we introduced ImagenHub to identify the true progress in the field of image synthesis and standardized the human evaluation protocol. then we proposed VIEScore to leverage the use of Multimodal LLMs to achieve human-like evaluation, yielding explainable metrics. In the second part focuses on video editing where we developed AnyV2V, a paradigm bridged the image and video editing problem with Image-to-Video models, setting up the stage to enhance the robust control in video manipulation.

Although the improvement has been substantial in the published papers, we noticed that most published works exhibit a high degree of inconsistency in their experimental setups. The existing studies create their own evaluation datasets, leading to a lack of comparability between different models. Some studies require extensive hyper-parameter tuning and prompt engineering to achieve acceptable performance, compromising the robustness of the model. The varying degrees of tuning effort across different models can result in unfair comparisons. Different human evaluation protocols and guidelines are used across studies, making it impossible to compare human evaluation scores. Additionally, some studies either rely on a single rater or do not report inter-rater agreement, further hindering the comparability of results between papers. These inconsistencies primarily stem from the evaluation datasets, inference methods, and evaluation criteria. Such a research environment is greatly hindering the development of the field because the true research progress and the most effective methods in the field cannot be truly identified.

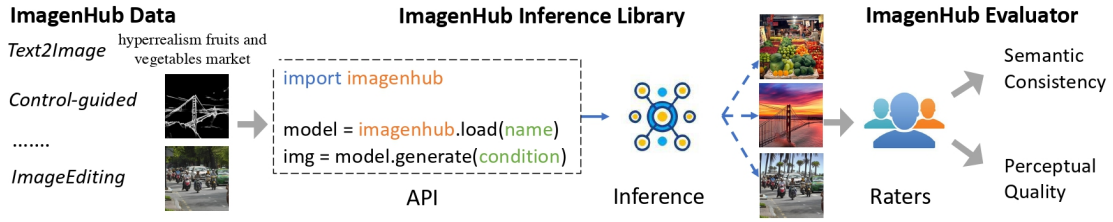


Figure 1.1: ImagenHub Data contains evaluation inputs. With the use of ImagenHub inference library, models generate outputs in a fair setting. The ImagenHub evaluators evaluate images according to the two carefully designed metrics.

To address the issue, we introduce ImagenHub, with an initiative to standardize the evaluation of conditional image generation models. ImagenHub contains 1) a diverse evaluation set that contains 7 subset for different conditional image synthesis tasks, each with 100-200 instances. 2) an inference library as a public tool to standardize the inference and evaluation of all the conditional image generation models. 3) designed human evaluation metrics: 'Semantic Consistency' and 'Perceptual Quality', which achieved high inter-worker agreement as measured by Fleiss' Kappa [29] and Krippendorff's Alpha [49]. Our rating standard offers several benefits: (1) It provides an unambiguous checklist table, making it easy for raters to evaluate images. (2) The rating guideline is consistent across all task types. (3) Sustainability is ensured, as previous evaluation results can be reused when new models are introduced since each model is rated individually, unlike preference-based ratings. Figure 1.1 illustrates the three main components in this part of the study. The ImagenHub framework is composed of three key elements: the newly curated ImagenHub dataset, the ImagenHub library, and the ImagenHub evaluation platform.

To complement ImagenHub, we developed VIEScore (Visual Instruction-guided Explainable Score), a metric that enables task awareness and the reasoning ability in evaluating sythesized images. It leverage multimodal Large Language Model (LLM) with a guiding prompt similar to ImagenHub human rating guidelines to rate images from human perspective, as illustrated in Figure 1.2. Our experiment showed although the current multimodal LLMs struggle in some minor issues, it's rating is highly align to human in Text-guided image generation tasks. VIEScore provides better evaluation than tradition metrics because 1) VIEScore provide not just the score but also the rationale. We want to enable the understanding of each judgment. 2) Traditional metrics are not task-aware, which means they cannot be broadly adapted to different tasks.

Moving to video, we look for solutions for precise video editing. The field of digital content creation using generative models is dynamic, but current state-of-the-art video

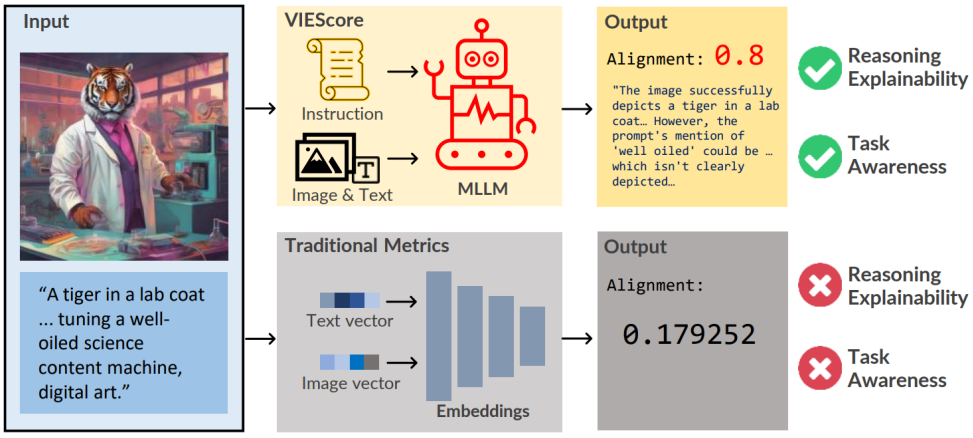


Figure 1.2: The difference between VIEScore and traditional metrics.

editing models still do not provide the level of quality and control that users desire. Previous video editing approaches either extended from image-based generative models in a zero-shot manner or required extensive fine-tuning, which can hinder the production of fluid video edits. Additionally, these methods often rely on textual input as the editing guidance, leading to ambiguities and limiting the types of edits they can perform.

To address these challenges, we introduce AnyV2V, a striaghtforward approach to tackle Video Editing problem, leveraging the research progress in image editing. AnyV2V simplifies the process into two main steps: (1) using an off-the-shelf image editing model to modify the first frame, and (2) utilizing an existing image-to-video generation model to generate the edited video through temporal feature injection, as shown in Figure 1.3.

The primary strength of AnyV2V lies in 1) its ability to utilize various existing image editing tools to facilitate a wide array of video editing tasks, such as prompt-based editing, reference-based style transfer, subject-driven editing, and identity manipulation—functions that were previously unattainable with older techniques. 2) its adaptive framework demonstrates enhanced performance with the integration of more robust Image-To-Video models, as supported by our evaluation findings across various backbone models.

By examining the image and video synthesis models, we not only identified the true progress in the field but also standardized and improved its evaluation. Then we also bridged the image editing problem and video editing problem. These contributions outline a path for their development towards models that are more robust at generation and manipulation.

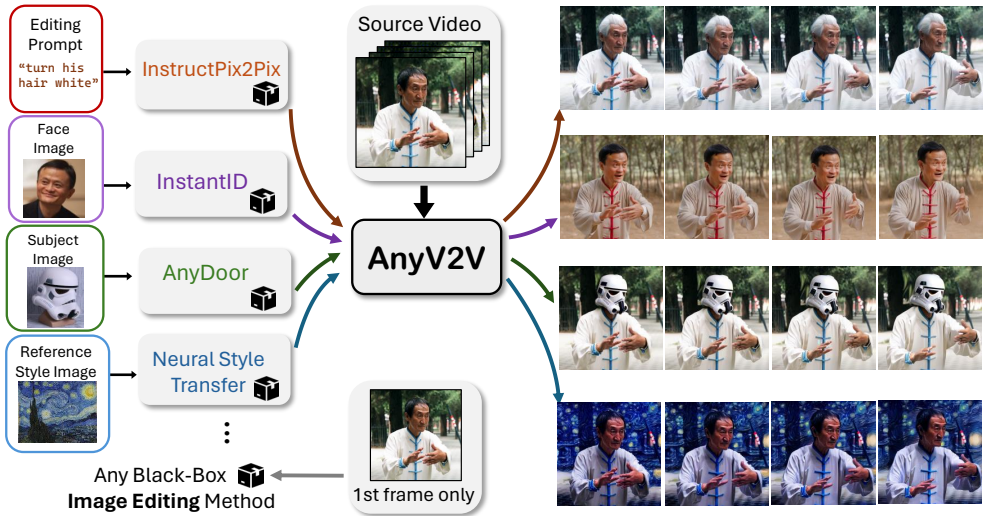


Figure 1.3: AnyV2V adapt image editing solutions into video editing.

Chapter 2

Background

This section surveys the methodologies of visual synthesis in the image and video domain. It contains pivotal studies related to conditional generation and manipulation in the image and video domain.

2.1 Image Generation and Editing

Image Generation is a popular topic since the development of GANs (Generative Adversarial Networks) [35]. It refers to synthesize an image that looks like it belongs to the distribution of our desired image. In recent years, diffusion models [42] have revolutionized the field due to its amazing properties in synthesizing high quality images. More recent works in the field focused on how to condition the synthesis through text [84, 68, 41], image [109, 77], and subject [85]. Most of the recent works relied on cross attention module in the diffusion model to incorporate conditional information into it [84]. These problems can be formulated as the following equation:

$$y = f(C) \quad \text{where} \quad C = [c_1, c_2, \ldots, c_k]$$

where y is the synthesized image, f is the image synthesis model (i.e. diffusion model) and C are the conditions which it can involve text or any kind of images.

Image Editing, refers to altering one or multiple parts of the image, can be either locally or globally. The edited image would preserve part of the semantic information from the original image. Researchers are also trying to provide different controls for the image

editing, such as instruction-guided [12, 1] and description-guided [8, 24]. In these problems, the original image became of the conditions in the synthesis process. We briefly introduce the common image generation and editing tasks that closely related to this work.

2.1.1 Text-Guided Image Generation

Text-Guided Image Generation, also known as Text-to-Image, is to generate an image consistent with the text description. It can be formulated as $y = f(p)$ where p is the prompt, as the description of the image. Pivotal works include Stable Diffusion [84] that trains the diffusion model in latent space, Imagen [41] employing a cascading diffusion model architecture, and PixArt-Alpha [18] using pixel-based diffusion transformer architecture. All of the methods leverage the cross attention in the transformer module and the CLIP [80]-like text encoder for aligning text and image modality.

2.1.2 Control-guided Image Generation

Control-guided Image Generation refers to the use of guidance image to control the generation process, first proposed in ControlNet [109], utilizing zero convolution, ControlNet allows the fine-tuning of Stable Diffusion models to gain image control. It can be formulated as $y = f(I_{\text{control}}, p)$ where I_{control} can be images like canny edges, sketches, pose extractions and even grayscale images.

2.1.3 Mask-guided Image Editing

Mask-guided image editing uses a mask to specify regions of an image for modification, allowing precise control over edits like object removal or attribute changes while preserving the rest of the image. It can be formulated as $y = f(p, I_{\text{src}}, I_{\text{mask}})$ where I_{src} is the source image and I_{mask} is a binary mask that one of color representing the to-be-edited image parts. Pivotal works include Glide [68], Blended Diffusion [3].

2.1.4 Text-Guided Image Editing

Similar to Mask-guided Image Editing, Text-Guided Image Editing perform image edit smartly with solely text prompt. The problem can be formulated as $y = f(p, I_{\text{src}})$, where I_{src} is the text prompt p can be a target description or an instruction prompt, depending

on the model used. Pivot works include InstructPix2Pix [12] and CosXL [1] leverage synthesized data edit pairs for training.

2.1.5 Subject-driven Image Generation

Subject-driven Image Generation, also known as image personalization, allows a customized item or concept to be appeared in the synthesized image, even if the model was not pre-trained on such data. This problem can be formulated as $y = f(S, p)$ where p is the prompt and S is a set of images regarding a specific subject, normally ranges from 3-5. Pivotal works include Textual Inversion [30] which optimize a text vector aligning to the S distribution, while DreamBooth [85] fine-tunes the whole model with class-specific prior preservation. BLIPDiffusion even allows S to be a single image.

Another new type of task, named multi-concept image composition, introduced in Custom Diffusion [54], generate images based on multiple subjects or concepts. The equation can be extended as $y = f(S_1, S_2, p)$ where S_1 and S_2 are the two different customized object or concept.

2.1.6 Subject-driven Image Editing

Subject-driven Image Editing refers to replacing or adding a subject in the scene that the final synthesized image would contain a customized object that perfectly blended in. This task was firstly proposed in DreamEdit [59]. The problem can be formulated as $y = f(S, p, I_{\text{src}})$ where p is the prompt, I_{src} is the input image and S is a set of images regarding a specific subject, normally ranges from 3-5.

2.2 Evaluation Metrics for Image Synthesis Tasks

2.2.1 Algorithm-based evaluation

Various metrics have been proposed to evaluate the quality of synthesized images. Traditional measures like the Inception Score (IS) [89] and the Frechet Inception Distance (FID) [40] are commonly used to assess image fidelity. To measure the alignment between generated images and text prompts, several metrics have been introduced [48, 55, 73, 88], with the CLIP score [39] and BLIP score [58] being the most widely used.

For image editing and control-guided image generation [64, 77], LPIPS [110] is frequently used. It measures the similarity between two images in a way that aligns with human perception. Its sensitivity to distortions allows it to compare the input (or ground truth) and the generated image. However, in the context of image editing, perceptual quality from a human perspective often includes factors like naturalness (e.g., shadows, lighting, sense of distance), which the LPIPS metric does not account for. Additionally, assessing a model's performance based on distortion levels is challenging, as current models often achieve high-quality editing without noticeable artifacts.

For subject-driven image generation and editing tasks, CLIP-I and DINO were often used [85]. CLIP-I calculates the average pairwise cosine similarities between CLIP embedding of generated and real images, as first introduced in Textual-Inversion [30]. However, CLIP-I struggles to differentiate between subjects with similar text descriptions and is less sensitive to shape consistency since it focuses on semantic similarity between images. DINO computes the mean cosine similarities between DINO embedding of ViT-S/16 [14] for both synthesized and real images. Unlike CLIP-I, the DINO metric is sensitive to differences between subjects of the same class due to its self-supervised training objective. These two metrics have become popular in research on subject-driven image generation and editing tasks [59].

2.2.2 Human-centered evaluation

Tradition method for evaluating synthesized images is to have human annotators to assess visual quality [47, 59, 24]. Most works conducted human preference study across baselines, making the results lack of insights on how good the models actually are. Recent works like HEIM [56] attempt to standardize human evaluation across various dimension, but scalability remains a challenge. The end goal is to develop an algorithm-based evaluator that highly align with human evaluations.

2.3 Video Generation and Editing

A Video is formed by a sequence of images. Similar to image synthesis problem, it can be formulated as:

$$V \in [I_1, I_2, \ldots, I_k] = f(C) \quad \text{where} \quad C = [c_1, c_2, \ldots, c_k]$$

where V is a video, I is an image, C are the conditions such as text or image. Most of the works relied on Diffusion models and attached a temporal module to give it ability to synthesize videos [100, 22, 7].

2.3.1 Text-To-Video (T2V) Generation

Also known as Text-guided Video Generation. Recent progress in diffusion models [42, 90] has significantly pushed forward the development of Text-To-Video (T2V) generation. Text-to-Video (T2V) models are a type of generative model that can synthesize new video sequences that didn't previously exist, based solely on textual prompts [100, 69, 16, 17]. Existing diffusion-based Text-to-Video (T2V) models generally build upon Text-To-Image (T2I) models, adding additional temporal modelling layers that understand how things change over time, thus extending it into video domain. Recent T2V generation models use advanced approaches based on diffusion processes, which can start from a random noise and gradually refine it into a detailed video that matches the text description [44, 41, 32]. Among these, models based on Latent Diffusion Models (LDMs) have gained particular attention for their effectiveness and efficiency [112, 2, 11, 37, 16, 100]. Some works also used the diffusion transformers [38, 19, 69]. LDM-based [84] Text-to-Video (T2V) generation models are the most extensively studied methods due to their high efficiency.

2.3.2 Image-To-Video (I2V) Generation

Also known as Image-guided Video Generation. In the realm of video creation from images, innovative methods have been developed to refine how videos are generated, focusing on incorporating specific images to guide the video's visual. Initial attempts like VideoComposer [99] and VideoCrafter1 [16] leveraged an image as a stylistic reference but struggled to maintain detailed visual consistency in their videos. However, newer approaches like DynamiCrafter [105], I2VGen-XL [111], SEINE [22], and ConsistI2V [83] have made significant improvement, producing videos that closely align with the visual cues of the input image as the starting frame. These Image-To-Video (I2V) generation models have found varied applications, including long video generation [22, 107, 83], generative frame interpolation [105, 22] to enhance fluidity and video storytelling [105, 107]. Despite their wide usage, there's a noticeable gap in employing these models for video editing tasks.

2.3.3 Video-To-Video (V2V) Generation

In the evolving landscape of digital media, Video-to-Video (V2V) Generation (also known as Text-guided Video Editing) is an approach that leverages generation models to perform video content manipulation. This innovative process allows creators to alter scenes, adapt styles, and even replace objects in a existing video. A Video-To-Video (V2V) model needs to have the capability to understand and manipulate the temporal flow and spatial elements of videos. Early works like Tune-A-Video [104] requires fine-tuning on a Text-To-Image (T2I) model to perform video editing, but also result in blurry and saturated videos. Later works [33, 102, 76, 23, 106, 15] also leveraged the Text-To-Image (T2I) model with improved performance, allowing users to perform intuitive video manipulation through textual prompts. Yet, these approaches do not provide users with exact control, leading to modifications that might not match the user's specific intentions or the desired detail level, due to the vagueness of natural language and the limitations of model capabilities. It is desired to deliver a comprehensive solution that ensures greater precision for video editing tasks.

Chapter 3

Identifying Effective Methods in Conditional Image Synthesis

The current research field of image synthesis is evolving at a fast pace. To identify the true research progress and the most effective methods in the field, we introduce ImagenHub, with an initiative to standardize the evaluation of conditional image generation models. In detail, we introduce ImagenHub data, a diverse evaluation set that contains 7 subset for different conditional image synthesis tasks, each with 100-200 instances. ImagenHub data ensures the fair comparisons across different models in the same task through standardizing the evaluation inputs. Then We also presented ImagenHub inference library as a public tool to standardize the inference and evaluation of all the conditional image generation models. We gathered and standardized the widely scattered codebases from existing works into a single, unified format. By fixing hyperparameters and prompt formats during inference, we eliminated the need for per-instance tuning, making the evaluation of different models both fair and reproducible. The library's design prioritizes ease of extension.

Finally, we investigated various human evaluation metrics to extract human perspective ratings and refined different versions of the rating guidelines to enhance inter-rater agreement. Ultimately, we adopted two three-valued rating metrics: 'Semantic Consistency' and 'Perceptual Quality', which achieved high inter-worker agreement as measured by Fleiss' Kappa [29] and Krippendorff's Alpha [49]. Most of our models achieve more than 0.4 on Krippendorff's Alpha [49]. Our rating standard offers several benefits: (1) It provides an unambiguous checklist table, making it easy for raters to evaluate images. (2) The rating guideline is consistent across all task types. (3) Sustainability is ensured, as previous evaluation results can be reused when new models are introduced since each model is rated individually, unlike preference-based ratings.

In this chapter, we will present the methodology, the image synthesis models evaluated, the model leaderboard for each task, and an exploratory study on the capabilities and limitations of the image synthesis models.

3.1 Methodology Design

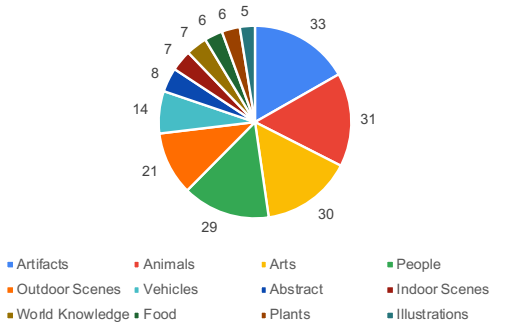
3.1.1 ImagenHub Data

We reviewed the existing public evaluation datasets [87, 101, 28, 108, 20, 54, 59] and selected a diverse range of instances to merge with our collected instances to create our ImagenHub dataset for seven generation tasks. The seven generation tasks include text-to-image generation, subject-driven image generation, multi-concept image composition (similar to subject-driven image generation but multiple subjects), control-guided image generation, mask-guided image editing, text-guided image editing, and subject-driven image editing. We have already discussed these tasks in the background chapter. We ensured the diversity in each subset of ImagenHub data, as illustrated in Figure 3.1.

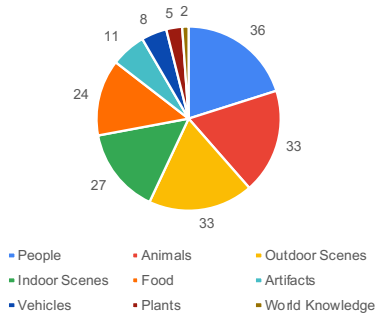
3.1.2 ImagenHub Inference Library

The ImagenHub inference library contains implementations of state-of-the-art image synthesis models that were originally scattered everywhere. All the models followed the official implementation or the settings recommended in the HuggingFace documentation [95]. The library provides an interface to generate outputs according to our ImagenHub data input. A visualization page of the benchmarking results is publicly available on the internet, thereby showcasing the true performance of each model, as shown in Appendix A.1 and A.2. By examining the actual results, viewers can gain valuable insights. Our framework also encourages future researchers to publicly release their human evaluation set generation results. By standardizing the human evaluation protocol, these results will become significantly more convincing due to increased experimental transparency.

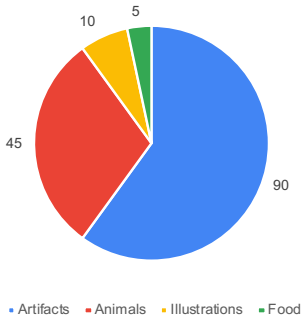
At the time of the thesis is written, the ImagenHub inference library contains the implementation of in total 52 conditional image synthesis models. In this part of study, 32 models are selected for evaluation. 7 models are categorized into Text-to-Image generation (text-guided image generation), 4 models in mask-guided image editing, 8 models in text-guided image editing, 5 models in subject-driven image generation, 3 models in subject-driven image editing, 3 models in multi-concept image composition and 2 models



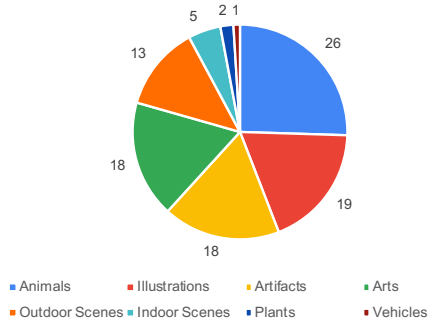
(a) Text-to-Image Generation dataset



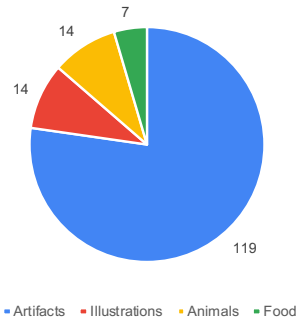
(b) Mask-guided/Text-guided Image Editing dataset



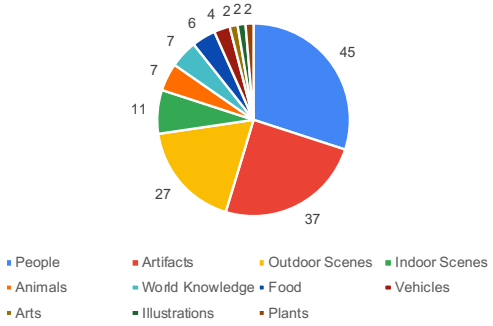
(c) Subject-driven Image Generation dataset



(d) Multi-concept Image Composition dataset



(e) Subject-driven Image Editing dataset



(f) Control-guided Image Generation dataset

Figure 3.1: Objects presented in each task's prompt text.

in control-guided image generation. We surveyed these models in Table 3.1. The number of parameters and runtime for Dalle-3 and MidJourney are not known. MidJourney is not open-source and lacks a whitepaper with technical details.

Model	#Params	Runtime	Keywords for technical detail
Text-to-Image Generation			
Dalle-3 [71]	-	-	Recaptioned training data
MidJourney [65]	-	-	-
Dalle-2 [81]	3.5B	10s	unCLIP, two-stage
Stable Diffusion [84]	0.8B	3s	Latent Diffusion
DeepFloydIF [26]	4.3B	37s	Cascaded Pixel Diffusion
OpenJourney [72]	0.8B	3s	SD, Midjourney data
Stable Diffusion XL [92]	2.3B	11s	Stable Diffusion, X-Large
Mask-guided Image Editing			
BlendedDiffusion [3]	0.5B	57s	Noise Blending, DDPM+CLIP
GLIDE [68]	3.5B	19s	CLIP, Diffusion
SD-Inpaint [86]	1.1B	11s	SD, Inpainting training
SDXL-Inpaint [92]	2.7B	36s	SDXL, Inpainting training
Text-guided Image Editing			
SDEdit [64]	1.3B	13s	SDE Prior
Text2Live [8]	3.1M	36s	Zero-shot, Edit layer
DiffEdit [24]	1.3B	29s	Mask estimation
Cycle Diffusion [103]	1.1B	9s	DPM-Encoder, Zero-shot
Prompt-to-Prompt [66]	1.1B	2m	Cross-Attention
Pix2PixZero [74]	1.1B	21s	Cross-Attention, Zero Prompt
InstructPix2Pix [12]	1.1B	11s	SD, synthetic P2P data
MagicBrush [108]	1.1B	7s	SD, MagicBrush data
Subject-driven Image Generation			
Textual Inversion [30]	1.1B	15m	Word embedding tuning
DreamBooth [85]	1.1B	10m	Finetuning with preservation loss
DreamBooth-Lora [46]	1.1B	8m	DreamBooth + Low-Rank Adaptation
SuTI [20]	2.5B	30s	In-context + Apprenticeship learning
BLIP-Diffusion [57]	1.1B	8s	Pretrained encoder, Zero-shot
Subject-driven Image Editing			
DreamEdit [59]	1.1B	8m	Dreambooth + Region proposal
PhotoSwap [36]	1.1B	7m	Dreambooth + Cross-Attention
BLIP-Diffusion [57]	1.1B	18s	Pretrained encoder, Zero-shot
Multi-concept Image Composition			
CustomDiffusion [54]	1.1B	19m	Cross-attention updating
DreamBooth [85]	1.1B	11m	Finetuning with preservation loss
TextualInversion [30]	1.1B	32m	Word embedding tuning
Control-guided Image Generation			
ControlNet [109]	1.4B	8s	Zero convolution + Frozen model
UniControl [77]	1.4B	23s	Multi-task pretraining, Zero-shot

Table 3.1: Evaluated models and their parameter size and runtime.

3.1.3 ImagenHub Evaluation

The human evaluation metric we designed can be applied uniformly across all seven tasks. We proposed two primary human evaluation metrics: Semantic Consistency (SC) and Perceptual Quality (PQ). These metrics assess the quality of the generated images from two different perspectives. Semantic consistency evaluates how well the generated RGB image y aligns with the condition $C = [c_1, c_2, \cdots, c_k]$, and each task has a different set of conditions. An overall score can obtained by computing the geometric mean of SC and PQ. The metrics can be formulated as the following equations:

$$SC(y, C) = min\{g_{SC}(y, c_1), g_{SC}(y, c_2), \cdots, g_{SC}(y, c_k)\} \in \text{List}[0, 0.5, 1] \tag{3.1}$$

$$PQ(y, C) = g_{PQ}(y) \in \text{List}[0, 0.5, 1] \tag{3.2}$$

$$O(y, C) = \sqrt{SC(y, C) \cdot PQ(y, C)} \in \text{Range}[0, 1] \tag{3.3}$$

where g_{SC} is the human rating process with the Semantic Consistency guideline and g_{PQ} is the human rating process with the Perceptual Quality guideline. We will discuss the two guidelines later. We choose min over mean for the SC formulation because it emphasizes the importance of meeting all criteria without exception. Using mean would also increase the complexity of human user rating and introduce a misalignment across tasks, as different tasks contain a different number of conditions c_i.

To ensure the reliability of the collected data, for each instance we introduced three human raters to rate the image. In that case. We can also compute the inter-worker agreement such as Fleiss Kappa [29] and Krippendorff's Alpha [49]. Moreover, the averaged score will have a more diverse range comparing to the three options list [0, 0.5, 1].

In implementation, we utilized the Prolific platform to collect human feedback for evaluating synthesized images. To ensure high-quality ratings, we set the following criteria for workers: 1) Workers must primarily speak English. 2) Workers must complete the rating of one set of images (approximately 150-200 images) in a single session. This ensures they fully understand the rating instructions and allows for consistent ratings by the same person, enabling us to conduct inter-worker agreement analysis across the entire benchmark. Workers will rate each image based on two criteria: SC and PQ, both on a scale of 0, 0.5, and 1. The instructions and images are hosted on a website, and workers are provided with detailed guidelines and examples to aid in the rating process. Responses are submitted in a TSV file. To ensure annotation quality, we pre-rated a few significant samples (around 10-20) for each task. These serve as references to assess annotator performance. Ratings that do not meet our standards will be rejected.

ImagenHub Human-rater Guideline

Human have to strictly follow the guide in the Table 3.2 when participating the rating process. The SC is measured by the coherency in terms of the condition provided (i.e. Prompts, Subject Token, etc.) in the synthesized image, while the PQ score is determined by the natural sense or the artifacts of the synthesized image. Artifacts includes distortions, watermarks, scratches, blurred faces, unusual body parts, or subjects not harmonized. While unusual sense referred to wrong sense of distance such as subject too big or too small compared to others, wrong shadow, or wrong lighting, etc.

Condition 1	Condition 2	Condition 3	SC rating
Inconsistent	Any	Any	0
Any	Inconsistent	Any	0
Any	Any	Inconsistent	0
Partially Consistent	Any	Any	0.5
Any	Partially Consistent	Any	0.5
Any	Any	Partially Consistent	0.5
Mostly Consistent	Mostly Consistent	Mostly Consistent	1.0

Subjects in image	Artifacts	Unusual sense	PQ rating
Unrecognizable	Any	Any	0
Any	Serious	Any	0
Recognizable	Moderate	Any	0.5
Recognizable	Any	Moderate	0.5
Recognizable	Little/None	Little/None	1.0

Table 3.2: Rating guideline for computing the SC and PQ score.

Ablation Study of Human Evaluation Design

We investigated the impact of the number of human raters on the evaluation of synthesized images. We recruited 10 raters from Prolific to assess 15 samples produced by the Midjourney model. As shown in Figure 3.2, the mean scores remained relatively stable with a slight increase at 4 raters and the highest mean score at 5 raters. The mean scores hovered around 0.70, while the error bars indicated an increase in standard deviation. Krippendorff's Alpha decreased as the number of raters increased. It started around 0.70 with 3 raters and dropped to 0.64 with 10 raters. This suggests that while the mean score

remains consistent, the reliability of that score decreases slightly with more raters due to increased variance. With this analyzed, we optimized our human evaluation protocol and used the setting of 3 raters, balancing the cost and reliability.

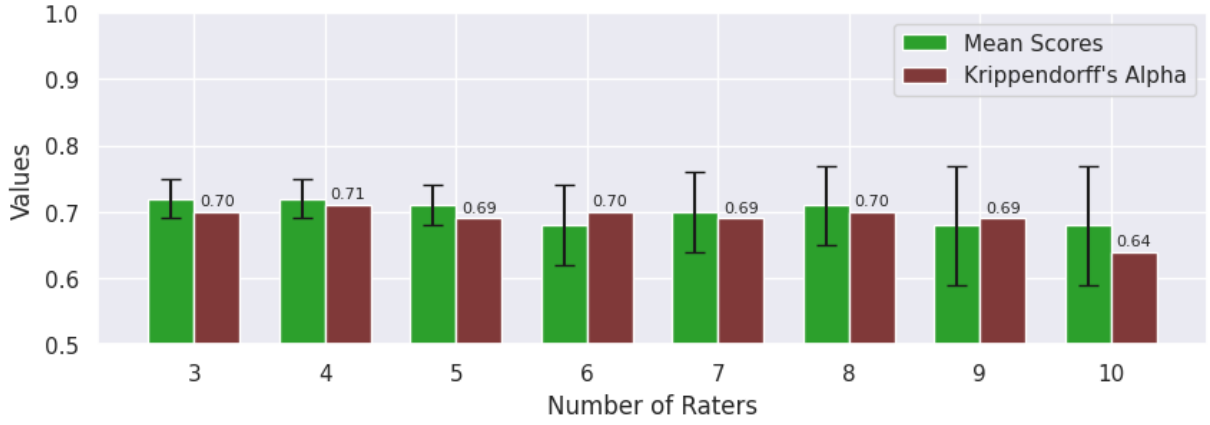


Figure 3.2: While the mean score stayed stable, the reliability measured by Krippendorff's Alpha diminished slightly as the number of raters increased, due to added variance.

In our study, we assumed the aspects would heavily impact each other in image rating, thus we derived that $O = \sqrt{SC \times PQ}$. But there could be other options such as a weighted importance $O = \alpha \times SC + \beta \times PQ$ to explore. We explored the settings through the image editing task in Table 3.3. Text2Live surpasses CycleDiffusion in the weighted sum setting, despite CycleDiffusion performing better in human evaluations. Our investigation revealed that Text2Live often returns the input as the output, resulting in SC=0 and PQ=1. We attempted to adjust the weightings, but this did not accurately represent the model's performance. Therefore, we opted to use the geometric mean setting to penalize the models more effectively.

Setting	$O = \sqrt{SC \times PQ}$		$O = 0.5SC + 0.5PQ$		$O = 0.7SC + 0.3PQ$	
	O_{Sum}	O_{Avg}	O_{Sum}	O_{Avg}	O_{Sum}	O_{Avg}
MagicBrush	83.51	0.47	103.75	0.58	98.85	0.55
CycleDiffusion	24.89	0.14	65.25	0.36	51.28	0.29
DiffEdit	1.71	0.01	22.08	0.12	14.38	0.08
Text2Live	4.08	0.02	75.25	0.42	46.75	0.26

Table 3.3: Comparison on overall human score computation setting.

Finally, we also studied the impact of the granularity of SC and PQ. We came up with three settings: scoring range [0, 1] for both SC and PQ scores led to overly polarized and extreme values. However, scoring range of [0, 0.5, 1, 2] for both SC and PQ, where 2 indicated a perfect image would resulted in suboptimal Fleiss Kappa values because its hard for a person to distinguish the score 1 and 2, as shown in Table 3.4. To achieve a balance between maintaining reasonable score ranges and ensuring reliable evaluation, we settled on a range of [0, 0.5, 1], accompanied by clear and explicit guidelines.

Setting	categories = [0, 0.5, 1]			categories = [0, 0.5, 1, 2]			categories = [0, 1]		
	O_{Avg}	$\text{Fleiss}_{\bar{O}}$	$\text{Kd}_{\bar{O}}$	O_{Avg}	$\text{Fleiss}_{\bar{O}}$	$\text{Kd}_{\bar{O}}$	O_{Avg}	$\text{Fleiss}_{\bar{O}}$	$\text{Kd}_{\bar{O}}$
BLIP-Diffusion	0.13	0.41	0.47	0.09	0.29	0.50	0.05	0.40	0.40
DreamEdit	0.50	0.33	0.52	0.32	0.26	0.55	0.21	0.38	0.38
PhotoSwap	0.62	0.35	0.46	0.36	0.25	0.50	0.29	0.37	0.37

Table 3.4: Ablation study to understand the impact of different evaluation settings.

3.2 Experiments Results

This section reports the results of the evaluation experiment. Over 14k of synthesized images are generated. During the image generation process we disabled negative prompts and any prompt engineering techniques for fairness. Then these images got sent to human raters. The human annotation took 2 weeks to complete. We hired 24 annotators. For the time used in evaluating one set of images (around 150-200 images), the mean time reported is 1 hour 40 minutes. Each of the annotators conducted evaluations on 3-4 sets (or more) of images. The human annotations data took estimated 150 hours to collect.

With the collected human data, we compiled a leaderboard in Table 3.5, with computed Krippendorff's Alpha [49] and Fleiss kappa [29]. Krippendorff's Alpha is expected to yield a higher value than Fleiss kappa. This distinction stems from the nature of the rating categories: Fleiss' Kappa assumes nominal categories, while Krippendorff's Alpha accommodates ordinal inputs. Both Fleiss' Kappa and Krippendorff's Alpha are bounded within the range of [-1, 1], where values greater than 0 indicate agreement, and values closer to 1 signify a higher degree of agreement. We can see that most of the models achieve a value higher than 0.3 on Fleiss kappa and a value higher than 0.4 on Krippendorff's Alpha, marking the reliability of the data.

Then we processed the data and found that the overall performance of the current models is generally unsatisfactory. From Figure 3.3 we can observe that the average overall

score for all tasks are below 0.4 except Text-guided Image Generation task. This indicate the field is still relatively immature. Looking into the detail, with Figure 3.4 we can see that most of the models can generate image with high perceptual quality while failing in semantic consistency, that is why most model are scattered in the top left corner of the figure. Only Text-guided Image Generation models achieved above 0.5 in both SC and PQ.

3.3 Insights from Human Inspection

To further look for insights, we manually inspected the synthesized image across all models for each task.

Inspection on Text-guided Image Generation (7 Models)

All models demonstrate the ability to generate high-quality images. They generally understand prompts well in terms of semantic consistency, but Stable Diffusion XL excels with more complex prompts. For instance, it accurately and intricately interprets "A panda making latte art," whereas other models often misinterpret this as "panda latte art." DALLE-3 and Midjourney also perform well with complex prompts, with DALLE-3 showing a slight edge in handling conflicting prompts.

Inspection on Mask-guided Image Editing (4 Models)

We observed that outputs from Stable Diffusion and GLIDE often contain noticeable artifacts around the masked region boundaries. While Blended Diffusion and Stable Diffusion XL avoid this issue, they frequently produce unrecognizable outputs. Although Stable Diffusion XL delivers the best results among these models, the overall performance is still unsatisfactory. Another common problem is that the filled regions struggle to blend seamlessly with the background.

Inspection on Text-guided Image Editing (8 Models)

We found that although Prompt-to-Prompt, Pix2PixZero, and SDEdit generate high-quality images, they often result in completely altered backgrounds. Additionally, in many instances, Text2Live simply returns the input as the output, a phenomenon occasionally

observed in other models as well. A crucial requirement is to edit the image precisely while keeping the background untouched. This task is particularly challenging because the network must accurately interpret the editing region from the semantic inputs. Our evaluation rankings align with the claims from CycleDiffusion, InstructPix2Pix, MagicBrush, Prompt-to-Prompt, and DiffEdit. However, our rankings do not align with Pix2PixZero, as their paper primarily tested word-swapping examples, which do not generalize well to more complex edits.

Inspection on Subject-driven Image Generation (5 Models)

Our evaluation results largely align with the findings of DreamBooth, BLIPDiffusion, and SuTI. Specifically, Textual Inversion struggles to maintain the features of target subjects. DreamBooth can imitate subjects based on provided images but sometimes ends up copying the learned images. DreamBooth-Lora has difficulty generating the desired subjects but can follow context prompts. BLIP-Diffusion can replicate target subject features but struggles with fine details. SuTI maintains high consistency with the chosen subjects and context, although it occasionally produces tolerable artifacts.

Inspection on Multi-concept Image Composition (3 Models)

Our evaluations confirm that CustomDiffusion consistently outperforms DreamBooth and TextualInversion. However, while it excels at learning the features of multiple given subjects, it often fails to follow prompts, particularly those involving actions and positional words. In contrast, DreamBooth successfully learns the correct subjects in some cases, whereas Textual Inversion rarely does. Interestingly, even when DreamBooth does not learn the correct subjects, it can still accurately follow the prompts in certain instances.

Inspection on Subject-driven Image Editing (3 Models)

It is crucial to modify the subject from the source to the target without significantly altering the background. PhotoSwap can naturally adapt the target subject from the source in most cases but rarely preserves the background well. DreamEdit maintains the context in most cases but sometimes leaves observable distortions at the edges of contextualization. BLIP-Diffusion often fails the adaptation, compromising for a more realistic generation. Our evaluation ranking does not align with the papers due to varying evaluation criteria.

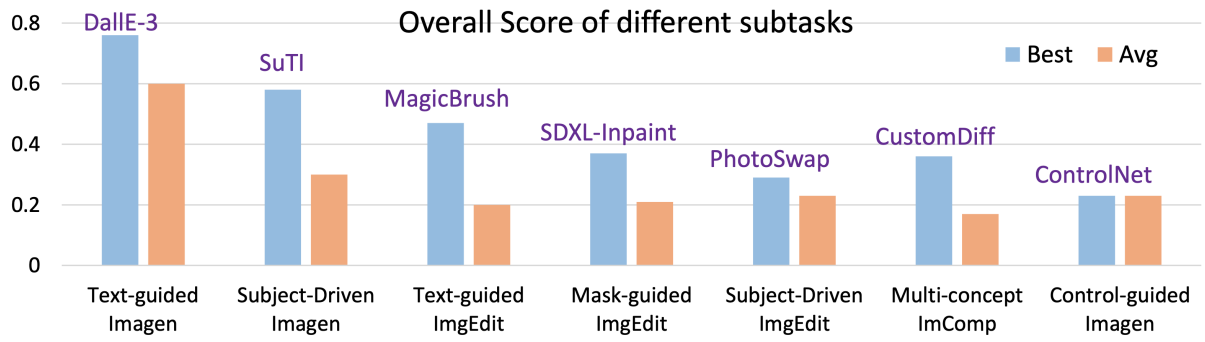


Figure 3.3: Overall score of the seven tasks in ImagenHub.

Inspection on Control-guided Image Generation (2 Models)

Our evaluation shows no significant difference between the two models in both automatic and human evaluation metrics. Although UniControl also reported no significant difference in automatic metrics, our human evaluation results do not align. This discrepancy may be due to different evaluation standards and aspects. Nevertheless, neither model demonstrates a high level of robustness, as scratches often appear on the generated images.

3.4 Key Takeaway

We proposed a library and a pipeline to standardize the human evaluation across image synthesis tasks. The study shows the true performance on different SOTA image synthesis models.

Through the manual inspection and the support from human-annotated data, The key takeaway is that 1) we believe that the development of Text-guided image generation has been quite mature. However, it is still difficult to apply complex control on the generation, given the performance of subject-driven and control-guided generation. 2) Image editing, despite being one of the most popular tasks, has not yet achieved the same similar performance as text-guide image generation. We concluded that we should focus more on the research on image editing and conditional image generation except for text condition.

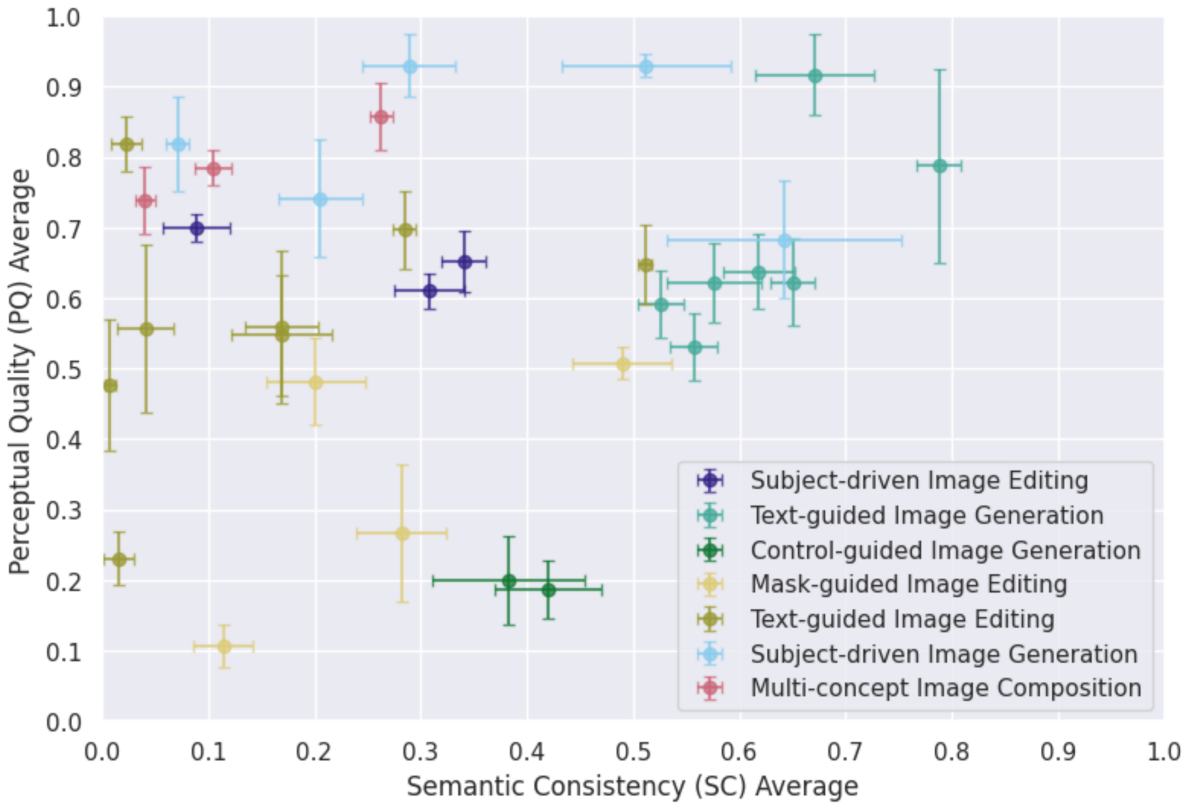


Figure 3.4: The SC and PQ score for each model in all categories.

Model	SC_{Avg}	PQ_{Avg}	Overall	$\text{Fleiss}_{\text{Overall}}$	$\text{Kd}_{\text{Overall}}$
Text-guided Image Generation					
Dalle-3	0.79±0.02	0.79±0.14	**0.76±0.08**	0.19	0.34
Midjourney	0.67±0.06	0.92±0.06	0.73±0.07	0.34	0.51
DeepFloydIF	0.65±0.02	0.62±0.06	0.59±0.02	0.32	0.51
Stable Diffusion XL	0.62±0.03	0.64±0.05	0.59±0.03	0.37	0.61
Dalle-2	0.58±0.04	0.62±0.06	0.54±0.04	0.27	0.40
OpenJourney	0.53±0.02	0.59±0.05	0.50±0.02	0.30	0.47
Stable Diffusion 2.1	0.56±0.02	0.53±0.05	0.50±0.03	0.38	0.50
Mask-guided Image Editing					
SDXL-Inpainting	0.49±0.05	0.51±0.02	**0.37±0.05**	0.50	0.72
SD-Inpainting	0.28±0.04	0.27±0.10	0.17±0.07	0.31	0.49
GLIDE	0.20±0.05	0.48±0.06	0.16±0.05	0.33	0.56
BlendedDiffusion	0.12±0.03	0.11±0.03	0.05±0.02	0.36	0.44
Text-guided Image Editing					
MagicBrush	0.51±0.01	0.65±0.06	**0.47±0.02**	0.44	0.67
InstructPix2Pix	0.29±0.01	0.70±0.06	0.27±0.02	0.55	0.74
Prompt-to-prompt	0.17±0.05	0.55±0.09	0.15±0.06	0.36	0.53
CycleDiffusion	0.17±0.03	0.56±0.11	0.14±0.04	0.41	0.63
SDEdit	0.04±0.03	0.56±0.12	0.04±0.03	0.13	0.13
Text2Live	0.02±0.01	0.82±0.04	0.02±0.02	0.10	0.17
DiffEdit	0.02±0.01	0.23±0.04	0.01±0.01	0.24	0.24
Pix2PixZero	0.01±0.00	0.48±0.09	0.01±0.01	0.37	0.37
Subject-driven Image Generation					
SuTI	0.64±0.11	0.68±0.08	**0.58±0.12**	0.20	0.39
DreamBooth	0.51±0.08	0.93±0.02	0.55±0.11	0.37	0.60
BLIP-Diffusion	0.29±0.04	0.93±0.04	0.35±0.06	0.22	0.39
TextualInversion	0.21±0.04	0.74±0.08	0.21±0.05	0.35	0.52
DreamBooth-Lora	0.07±0.01	0.82±0.07	0.09±0.01	0.29	0.37
Subject-driven Image Editing					
PhotoSwap	0.34±0.02	0.65±0.04	**0.36±0.02**	0.35	0.46
DreamEdit	0.31±0.03	0.61±0.03	0.32±0.03	0.33	0.52
BLIP-Diffusion	0.09±0.03	0.70±0.02	0.09±0.03	0.41	0.47
Multi-concept Image Composition					
CustomDiffusion	0.26±0.01	0.86±0.05	**0.29±0.01**	0.73	0.88
DreamBooth	0.11±0.02	0.78±0.02	0.13±0.02	0.61	0.71
TextualInversion	0.04±0.01	0.74±0.05	0.05±0.01	0.62	0.77
Control-guided Image Generation					
ControlNet	0.42±0.05	0.19±0.04	**0.23±0.04**	0.37	0.57
UniControl	0.38±0.07	0.20±0.06	0.23±0.07	0.36	0.58

Table 3.5: ImagenHub leaderboard for seven core tasks.

Chapter 4

Explainable Metric for Evaluating Synthesized Image

After standardizing the evaluation of all the conditional image generation models and identifying the capabilities of the current models, we hoped to continue our efforts to evaluate newly published models and update our leaderboard to keep track of the progress. However, we noticed the human-centered evaluation from ImagenHub is economically expensive to scale up. It would be desirable to come up with a machine-driven evaluator that can replace human evaluation while holding a similar perspective as human judgment. As a result, we took the initiative and proposed the use of a multimodal Large Language Model (LLM) for evaluating synthesized images. The multimodal LLM is fed with a guiding prompt similar to ImagenHub human rating guidelines to rate images from human perspective. Named as VIEScore (Visual Instruction-guided Explainable Score), this method enables task awareness and the reasoning ability that traditional metrics lack, as illustrated in Figure 1.2. We believed the metrics in the future would provide not just the score but also the rationale. We want to enable the understanding of each judgment. We also noticed that traditional metrics are not task-aware, which means they cannot be broadly adapted to different tasks.

We designed a benchmark that compares the evaluation from ImagenHub human raters and different available multimodal LLMs. At the time of this thesis was written, GPT4-o achieved a 0.4 Spearman correlation with human evaluations, while the human-to-human correlation was only 0.45. We also studied the correlation between traditional metrics and human annotation data, in comparison to our proposed VIEScore. Results show that our method beats traditional metrics in terms of correlations across all image synthesis tasks.

In this chapter, we will present the methodology, the models evaluated, the evaluation results, and a discussion on the challenges of using multimodal LLM for producing explainable metrics.

4.1 Methodology Design

4.1.1 Method of VIEScore

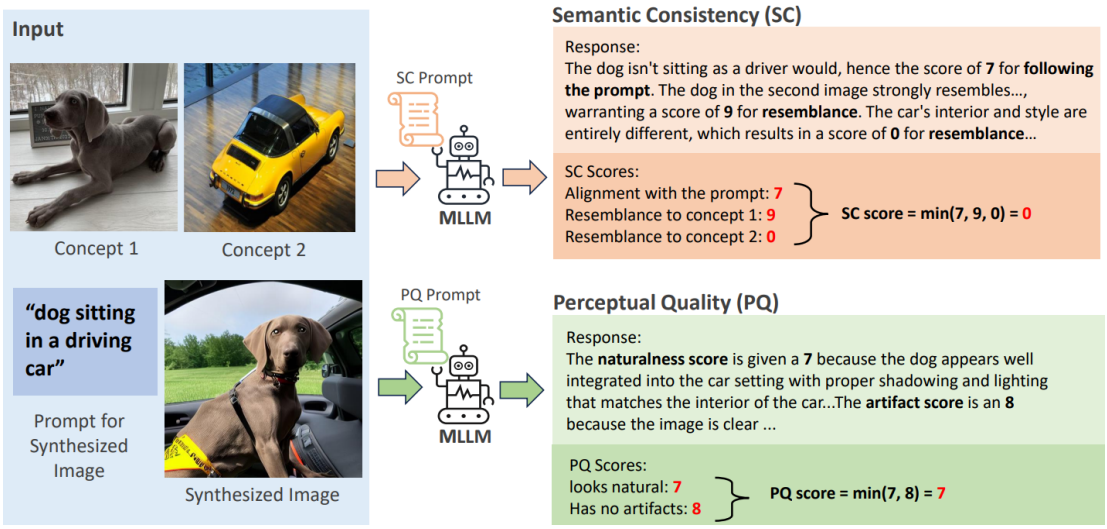


Figure 4.1: VIEScore takes all the conditions and the synthesized image as input. Then the input is fed to the MLLM with the instruction prompts, producing desired metrics aligned with the instruction prompt.

VIEScore takes the input as a set that contains the synthesized image y and the conditions $C = [c_1, c_2, \cdots, c_k]$. Then the multimodal LLM (MLLM) is fed with the input and the instruction prompt which define the desired metric requirement we want. The details of the instruction prompt can be found in Appendix B.1. VIEScore also produces two human perspective metrics, semantic consistency (SC) and perceptual quality (PQ), similar to ImagenHub. While each image in one rating aspect is rated by picking an option from List[0, 0.5, 1] by three human raters in ImagenHub, we propose a more rigorous rating instruction toward comprehensive evaluation for each type of task. We split the rating of semantic consistency (SC) and perceptual quality (PQ) into multiple sub-scores, which SC

contains multiple scores according to the task conditions. The score will be in the range [0, 10], The rationale behind the design is that in ImagenHub's human rating system, when the ratings from three human raters (each selecting from List[0, 0.5, 1]) are combined and averaged, the results will be one of the following: List[0.0, 0.17, 0.33, 0.5, 0.67, 0.83, 1.0]. Therefore, we use a 0 to 10 scale, normalized to the range [0.0, 1.0], for comparison with human ratings. Once the SC and PQ scores are retrieved, the overall score is derived as the geometric mean of both SC and PQ, as the same implementation in ImagenHub.

4.1.2 Experiment Setup

To comprehensively verify this method, we conduct a study on a wide range of image synthesis tasks to compare the correlations. We selected a list of available Multimodal LLMs as the backbone of our VIEScore. The list included GPT-4o [70], GPT-4v [70], Gemini-Pro-1.5 [82], LLaVA [60], Qwen-VL [6], BLIP2 [25], InstructBLIP [25], Fuyu [9], CogVLM [98], and OpenFlamino [4]. We run VIEScore on the ImagenHub data from 29 image generation models across 7 tasks, which contain over 14k sets of conditions and synthesized images. Then we compare the Spearman correlations of VIEScore on different backbones and the human ratings from ImagenHub, as illustrated in Figure 4.2. The goal of this part of the study is to find out whether, at the current state, we can replace traditional metrics with our proposed VIEScore on image evaluation tasks. We will compare the Spearman correlations against the traditional metrics, and also the human raters.

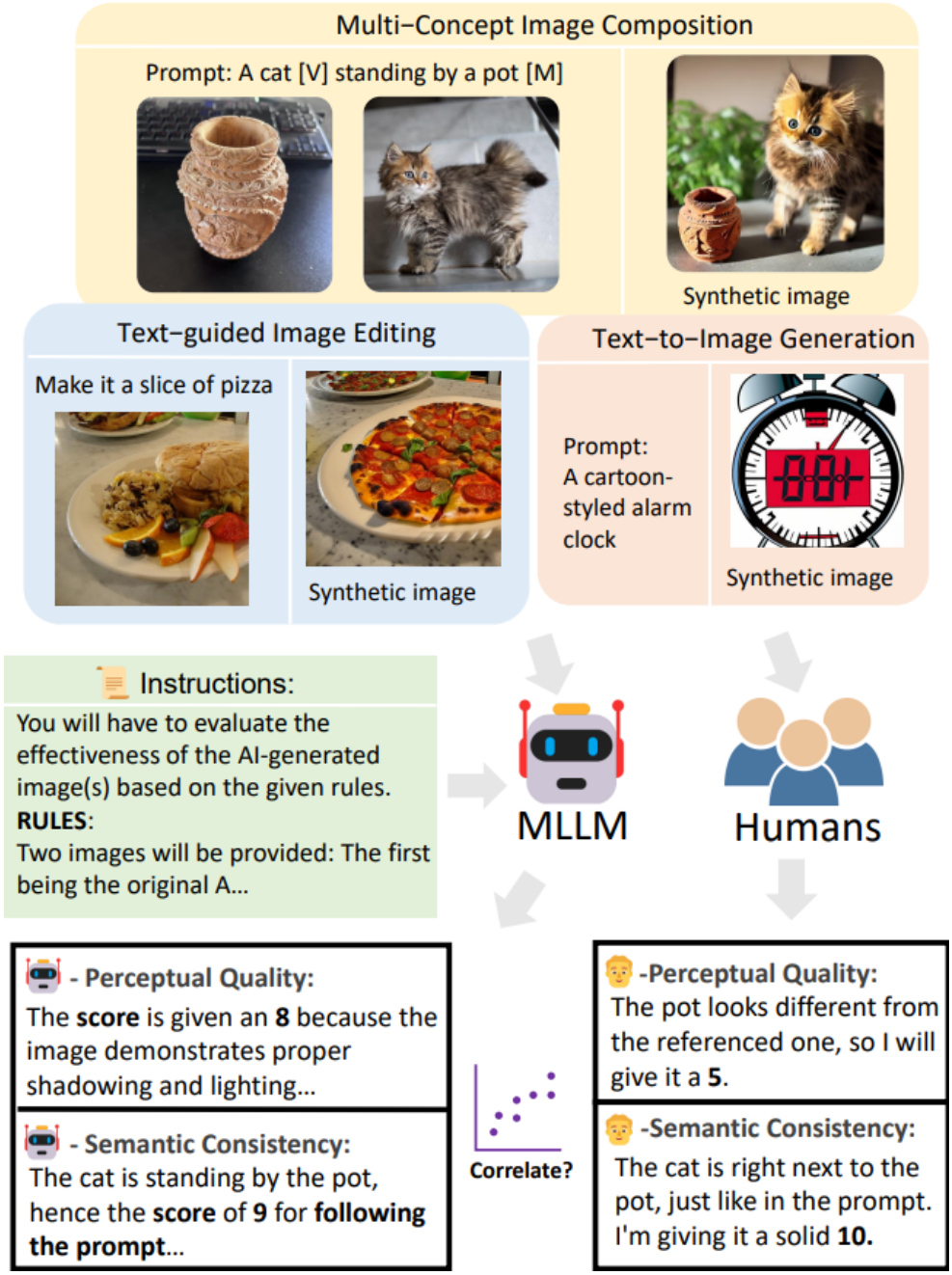


Figure 4.2: The experiment study the Spearman correlation between MLLMs and human perspectives on rating images across all tasks in ImagenHub.

4.2 Experiment Results

We applied Fisher Z-transformation to estimate the average Spearman correlation $\in [-1, 1]$ across models and tasks for all presented correlations. We first derived the baseline correlation by computing the Human-to-Human (H-H) correlation, which averaged around 0.5 and was expected to be the highest compared to MLLMs. Then we run VIEScore with different MLLM backbones. During the experiment, we observed that only GPT-4v, GPT-4o, Gemini, and LLaVA were able to clearly follow our instructions, whereas other MLLMs failed to produce meaningful results under our setup. For instance, BLIP-2, despite outputting the correct format, provided scores that were consistently zero. Qwen-VL and InstructBLIP managed to generate partial responses for semantic consistency but failed to produce any results for perceptual quality evaluation. We filled the meaningless results with random numbers.

4.2.1 Correlation Study against different backbones

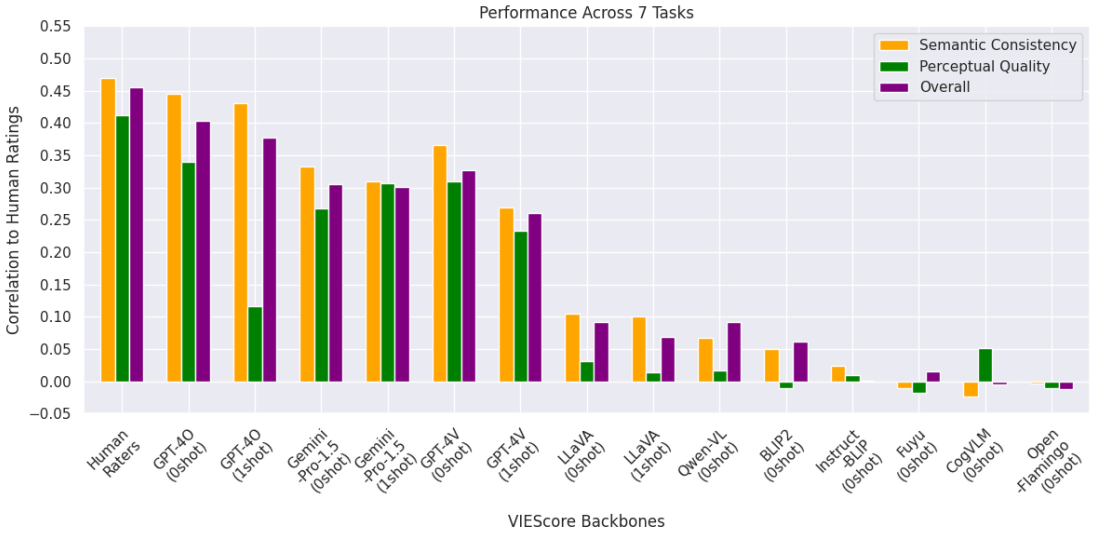


Figure 4.3: Backbone comparison of VIEScore in terms of correlation.

In terms of overall performance, GPT-4o exhibited a significantly higher correlation than all other MLLMs, while LLaVA's correlation was much lower compared to human raters. This suggests that LLaVA is less effective in these specific tasks compared to closed-source

MLLMs like GPT-4v. We also noticed the performance of closed-source MLLMs is generally much better than the open-source MLLMs, marking a big gap in this ability. Notably, GPT-4o achieved a very high correlation with human raters across various image generation and editing tasks, achieving 0.4 on the overall metric. GPT-4v and Gemini also performed well on nearly all tasks, with a human correlation of around 0.3. However, GPT-4v, Gemini, and LLaVA showed the weakest performance in text-guided editing and subject-driven image editing tasks. The detailed correlation numbers can be found in Appendix B.2, and we found that GPT-4v matches human performance in text-guided image generation tasks.

Numerous studies [13, 75, 61] have emphasized that In-Context Learning (ICL) enables LLMs to effectively handle novel tasks without traditional fine-tuning. We applied In-Context Learning in our prompting method, i.e. in 1 shot setting, anticipating an increase in correlation scores. However, we observed the opposite effect. there is a general trend of diminishing correlation scores except for Gemini-Pro-1.5.

4.2.2 Correlation Study against traditional metrics

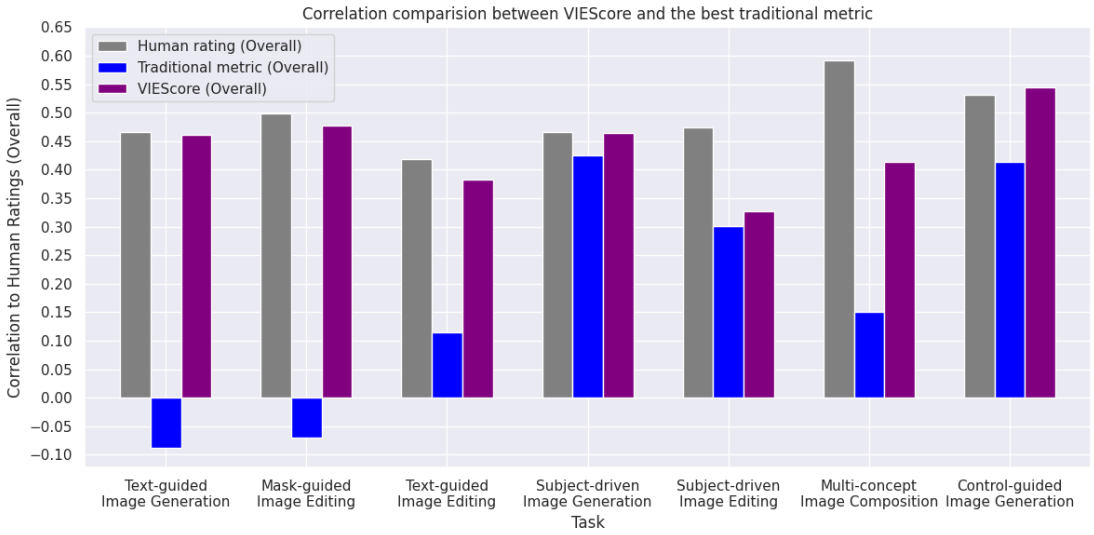


Figure 4.4: Comparison of VIEScore and traditional in terms of correlation by each task.

We explored the correlation to humans based on VIEScore and Traditional metrics across all 7 image synthesis tasks from ImagenHub. We noticed VIEScore beats Traditional metrics in every task, such as the CLIPScore [39] from the Text-guided image generation

Task (Total number of Models)	$d_{SF}(r_{\text{Human}}, r_{\text{Method}})\downarrow$				$\rho_S(r_{\text{Human}}, r_{\text{Method}})\uparrow$			
	GPT-4v	LLaVA	LPIPS	CLIP	GPT-4v	LLaVA	LPIPS	CLIP
Text-guided Image Generation (5)	2	6	N/A	8	0.90	0.50	N/A	-0.20
Mask-guided Image Editing (4)	2	8	2	0	0.80	-1.00	0.80	1.00
Text-guided Image Editing (8)	12	16	20	16	0.67	0.48	0.17	0.48
Subject-driven Image Generation (4)	4	6	0	6	0.20	-0.40	1.00	-0.20
Subject-driven Image Editing (3)	2	2	4	4	0.50	0.50	-0.50	-1.00
Multi-concept Image Composition (3)	0	0	2	2	1.00	1.00	0.50	0.50
Control-guided Image Generation (2)	0	0	0	2	1.00	1.00	1.00	-1.00

Table 4.1: Ranking judgment from each metric method. $d_{SF}(r, r_*)$ is the Spearman's footrule and $\rho_S(r, r_*)$ is the Spearman's rho (correlation). LPIPS is not available for the first task because there are no reference images.

task, LPIPS [110] from the image editing tasks, and DINO [85] from the subject-driven image synthesis tasks. The detailed correlation numbers can be found in Appendix B.2, and we found that DINO from the subject-driven image generation task has a very high correlation with human ratings.

We also computed the correlation based on the ImagenHub rankings and the rankings predicted by VIEScore and traditional metrics, as shown in Table 4.1. We selected GPT-4v and LLaVA as VIEScore representatives because they are the most significant closed-source and open-source MLLMs. For traditional metrics, LPIPS and CLIP Score are selected. We computed Spearman's footrule $d_{SF}(r, r_*) \in [0, +\infty)$ and Spearman's rho $\rho_S(r, r_*) \in [-1, 1]$ to examine the ranking correlation. Both GPT-4v and LLaVA were able to align with the ImagenHub rankings on the multi-concept image composition task and control-guided image generation task, with only one model difference in the subject-driven image editing task. Although the results varied significantly across other tasks, GPT-4v generally showed a stronger alignment with the ImagenHub rankings compared to LLaVA.

4.3 Insights and Challenges

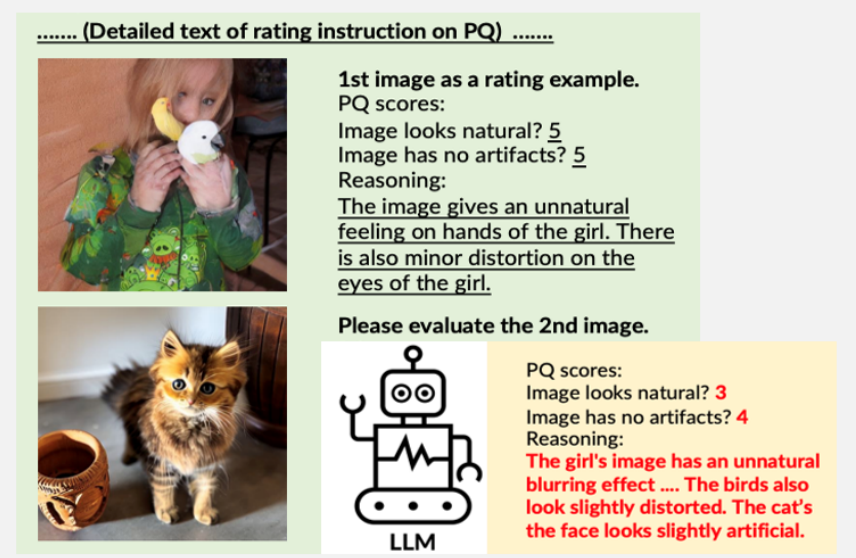


Figure 4.5: MLLM confused by the example image when rating images.

We looked into the reasonings and realized that the general MLLMs struggle in in-context learning when multiple images exist. The model will get confused. For example, a girl image is fed as an example to teach the model how to evaluate the cat image. Then the model starts rating based on the girl image and also the cat image. This is likely related to the poor ability of MLLMs to understand multiple images as illustrated in Figure 4.5. This behavior was observed in GPT-4v, GPT-4o, Gemini, and LLaVA. Another related work [62] also reported a similar issue, where the model attempted to consider the example when answering visual questions. This explains the deterioration of correlation scores for GPT-4v, GPT-4o, Gemini, and LLaVA when using the ICL prompting technique.

Moreover, One big issue that we found is that MLLMs often struggle to see the minor differences in two similar content images, as MLLMs tend to understand the images at the structure level but not detail enough, as illustrated in Figure 4.6. This suggests that the low correlation scores on image editing tasks are due to the limited capability of state-of-the-art MLLMs to understand multiple images.

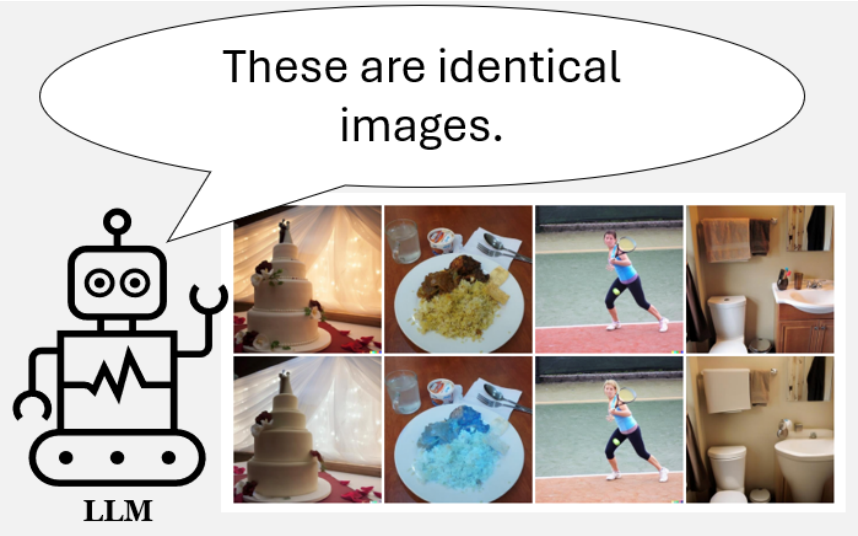


Figure 4.6: MLLM failed to distinguish minor differences in the images.

4.4 Key Takeaway

We are still far away from achieving explainable metrics for conditional image synthesis tasks. There are still obstacles for us to deploy explainable metrics in the current state. The current MLLMs cannot understand multiple images in the correct context and MLLMs often struggle to discern minor differences between two similar content images. However, there is still room to improve, and we believe VIEScore will be a mainstream evaluation method shortly, as in practice multiple image QA is an area coming with many applications and the MLLMs will certainly be improved in the future.

Chapter 5

Bringing Image Editing into Video Editing

In the second part of the study, we look for solutions for precise video editing. The field of digital content creation using generative models is dynamic, but current state-of-the-art video editing models still do not provide the level of quality and control that users desire. Previous video editing approaches either extended from image-based generative models in a zero-shot manner or required extensive fine-tuning, which can hinder the production of fluid video edits. Additionally, these methods often rely on textual input as the editing guidance, leading to ambiguities and limiting the types of edits they can perform. To address these challenges, we introduce AnyV2V, a striaghtforward approach to tackle Video Editing problem. AnyV2V simplifies the process into two main steps: (1) using an off-the-shelf image editing model to modify the first frame, and (2) utilizing an existing image-to-video generation model to generate the edited video through temporal feature injection.

The key advantage of AnyV2V is that it can leverage any existing image editing tools to support a wide range of video editing tasks, including prompt-based editing, reference-based style transfer, subject-driven editing, and identity manipulation - tasks that were not achievable with previous methods. Moreover, its evolving paradigm allows better performance when a stronger I2V model is applied, which has been back with our evaluation results on different backbone models.

Finally, our evaluation results show that AnyV2V significantly outperforms other baseline methods in both automatic and human evaluations. It maintains visual consistency with the source video while achieving high-quality edits across all the editing tasks.

In this chapter, we will present the methodology, task applications, and evaluation results. Then we include ablation studies and discussions on each component.

5.1 Methodology Design

Overview. The AnyV2V framework offers a streamlined approach to video editing, beginning with processing a source video. Initially, an image editing model is applied to edit the first frame of the video, customized for the specific editing requirement. This editing can be carried out either by a black-box image editing model ([12, 97, 21, 31]) or traditional editing software like photoshop. At the same time, the source video is transformed back into a noise latent state through an inversion process using DDIM sampling. During the denoising phase, spatial features, self-attention features, and temporal attention features are infused into the decoder layers of the I2V model. The inverted latent serves as the initial state in the denoising process, with the edited first frame used as the control signal. Each of the extracted features and attention mechanisms is carefully integrated into the appropriate decoder layers throughout the denoising stage. This method is highly controllable because it brings video editing into image editing problem. Figure 5.1 shows the complete pipeline of AnyV2V, while Figure 1.3 shows the example use cases of AnyV2V. We explore the design of each component in this section.

5.1.1 Flexible First Frame Editing

In visual manipulation, controllability is crucial for precise editing. AnyV2V enhances controllable video editing by leveraging image editing models to modify the video's first frame. This method allows for highly accurate modifications in the video and is compatible with a wide range of image editing models. These include deep learning models for image style transfer [31, 34, 63, 96], mask-based image editing [68, 3], image inpainting [93, 53], identity-preserving image editing [97], and subject-driven image editing [21, 59, 36].

5.1.2 Structural Guidance using DDIM Inverison

To ensure that the videos generated by the I2V generation model adhere to the overall structure of the source video, we use DDIM inversion to extract the latent noise of the source video at each time step t. This inversion is performed *without* a text prompt

condition but *with* the first frame condition. Given a source video $V^S = \{I_1, I_2, I_3, ..., I_n\}$, we obtain the inverted latent noise for time step t as:

$$\mathbf{z}_t^S = \text{DDIM_Inv}(\epsilon_\theta(\mathbf{z}_{t+1}, I_1, \varnothing, t)), \tag{5.1}$$

where DDIM_Inv$(\cdot)$ denotes the DDIM inversion process as described in Appendix C.1. The latent noise $\mathbf{z}_T^S$ at the final time step T (initial noise of the source video) should be used as the initial noise for sampling the edited videos in ideal situation.

5.1.3 Appearance Guidance via Spatial Feature Injection

Our observation indicates that I2V generation models possess inherent editing capabilities when utilizing only the edited first frame and DDIM inverted noise as inputs. However, this straightforward method often struggles to 1) Correctly preserve the background in the edited first frame and 2) Maintain the motion from the source video. This is primarily because the conditional signal from the source video encoded in the inverted noise is limited.

We perform feature injection in both convolution layers and spatial attention layers in the denoising U-Net to enforce consistency with the source video. In the video sampling process, We simultaneously denoise the source video using the previously collected DDIM inverted latents $\mathbf{z}_t^S$ at each time step t such that $\mathbf{z}_{t-1}^S = \epsilon_\theta(\mathbf{z}_t^S, I_1, \varnothing, t)$. We preserve two types of features during source video denoising: 1) convolution features f^{l_1} before skip connection from the l_1^{th} residual block in the U-Net decoder, and 2) the spatial self-attention scores $\{A_s^{l_2}\}$ from $l_2 = \{l_{low}, l_{low+1}, ..., l_{high}\}$ layers. We collect the queries $\{Q_s^{l_2}\}$ and keys $\{K_s^{l_2}\}$ instead of directly collecting $A_s^{l_2}$ as the attention score matrices are parameterized by the query and key vectors. We then replace the corresponding features during denoising the edited video in both the normal denoising branch and the negative prompt branch for classifier-free guidance [43]. We use two thresholds τ_{conv} and τ_{sa} to control the convolution and spatial attention injection to only happen in the first τ_{conv} and τ_{sa} steps during video sampling.

5.1.4 Motion Guidance through Temporal Feature Injection

We observe that the edited videos still frequently exhibit incorrect motion compared to the source video. On the other hand, we notice that video diffusion models in general, are often initialized from pre-trained T2I (Text-to-Image) models and then further trained on video data. During this training process, the parameters in the spatial layers are typically frozen

or set to a lower learning rate to preserve the pre-trained weights from the T2I model, while the parameters in the temporal layers are more extensively updated. Consequently, a significant portion of the motion information is likely encoded in the temporal layers of the I2V generation models.

Concurrent work [5] also observes that features in the temporal layers exhibit characteristics similar to optical flow [45], a pattern commonly used to describe the motion within a video. To better reconstruct the source video motion in the edited video, we propose to also inject the temporal attention features in the video generation process. Similar to spatial attention injection, we extract the source video temporal self-attention queries $Q_t^{l_3}$ and keys $K_t^{l_3}$ from some U-Net decoder layers represented by l_3 and inject them into the edited video denoising branches. We also only apply temporal attention injection in the first τ_{ta} steps during sampling.

5.1.5 Overall Formulation

Combining the spatial and temporal feature injection mechanisms, we replace the generating features $\{f^{*l_1}, Q_s^{*l_2}, K_s^{*l_2}, Q_t^{*l_3}, K_t^{*l_3}\}$ with the features from the source:

$$\mathbf{z}_{t-1}^* = \epsilon_\theta(\mathbf{z}_t^*, I^*, \mathbf{s}^*, t\ ; \{f^{l_1}, Q_s^{l_2}, K_s^{l_2}, Q_t^{l_3}, K_t^{l_3}\}), \tag{5.2}$$

where $\epsilon_\theta(\cdot\ ; \{f^{l_1}, Q_s^{l_2}, K_s^{l_2}, Q_t^{l_3}, K_t^{l_3}\})$ denotes the feature replacement operation across different layers l_1, l_2, l_3. This allows us to perform video editing without any finetuning.

5.1.6 Experiment Setup

Since such framework allows the use of I2V models, we experiment AnyV2V with three different I2V backbones, including SEINE [22], ConsistI2V [83], and I2VGenXL [111]. We also experiment different tasks with image editing models [21, 97, 12, 31], according to the 4 designed tasks applications:

1. **Prompt-based Editing:** allows users to manipulate video content using natural language alone. By providing descriptive prompts or specific instructions, users can easily perform a wide range of edits, such as adding accessories, swapping or spawning objects, incorporating effects, or changing the background.

2. **Reference-based Style Transfer:** In style transfer tasks, the artistic styles of Monet and Van Gogh are commonly investigated. However, in practical scenarios, users may wish to apply a unique style derived from a specific piece of artwork. Reference-based style transfer is one approach to achieve this, where a style image is used as a reference to edit videos. The goal is for the edited video to reflect the unique style of the reference artwork.

3. **Subject-driven Editing:** In subject-driven video editing, the objective is to replace an object in the video with a target subject derived from a given subject image. This process should preserve the original motion of the video and maintain the background intact.

4. **Identity Manipulation:** Identity manipulation enables users to alter video content by substituting a person in the video with another individual's identity, based on an input image of the target person.

To verify the robustness of this paradigm, we computed CLIP-based metrics and conducted human evaluation study. We compared

For the video generated in this experiement using AnyV2V, we used $0.2T$ for convolution features injection, $0.2T$ for the self-attention features injection, and $0.5T$ for the temporal attention features injection. T is the total number of sampling steps according to the DDIM sampler. We found that setting $0.2T$ for both convolution features and self-attention features injection is the best to preserve the layout and structure in the modified video without incorporating unnecessary visual elements from the original video. I also found that $0.5T$ is an optimal setting for temporal attention feature injection, as it provides a balance between motion alignment, motion consistency, and the overall quality of the video. We explain the design choices in the ablation study Section 5.2.2.

5.2 Experiment Results

We conducted all experiments on a single A6000 GPU, which can inference AnyV2V in 100 seconds. AnyV2V only takes 15000 GPU memory for the whole process.

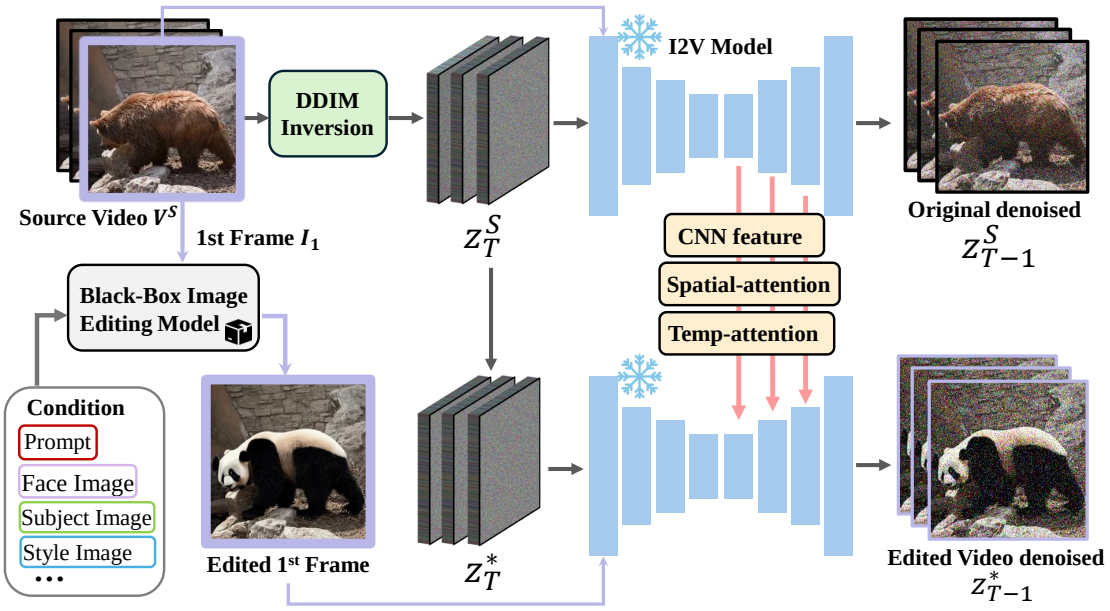


Figure 5.1: AnyV2V pipeline.

Table 5.1: Quantitative comparisons for AnyV2V with baselines on prompt-based video editing. Alignment: prompt alignment; Overall: overall preference. **Bold**: best results; Underline: top-2.

Method	Human Evaluation ↑		CLIP Scores ↑	
	Alignment	Overall	CLIP-Text	CLIP-Image
Tune-A-Video	15.2%	2.1%	0.2902	0.9704
TokenFlow	31.7%	20.7%	0.2858	**0.9783**
FLATTEN	25.5%	16.6%	0.2742	0.9739
AnyV2V(SEINE)	28.9%	8.3%	0.2910	0.9631
AnyV2V(ConsistI2V)	33.8%	11.7%	0.2896	0.9556
AnyV2V(I2VGen-XL)	**69.7%**	**46.2%**	**0.2932**	0.9652

5.2.1 Evaluation across Applications

Prompt-based Editing

Modifying video attributes through natural language commands, such as adding accessories, exchanging objects, or changing backgrounds, is a challenging and intricate en-

deavor. Unlike other advanced methods that often introduce unwarranted changes not specified in the text commands, the AnyV2V framework utilizes the precision of image editing models to ensure only the targeted areas of the scene are altered, leaving the rest unchanged. Combining AnyV2V with the instruction-guided image editing model InstructPix2Pix [12], AnyV2V accurately places a party hat on an elderly man's head and transforms a horse into a unicorn. Additionally, it maintains the original video's background and integrity, whereas in comparison, baseline methods often alter the color tone and shape of objects, as illustrated in Figure 5.2.

We conduct a human evaluation to examine the degree of prompt alignment and overall preference of the edited videos based on user voting, and also compute the metrics CLIP-Text for text alignment and CLIP-Image for temporal consistency. In detail, CLIP-Text is computed by the average cosine similarity between text embeddings from CLIP model [79] and CLIP-Image is computed in the same way but between image embeddings for every pair of consecutive frames. Table 5.1 shows that AnyV2V generally achieves high text-alignment and temporal consistency, while AnyV2V with I2VGen-XL backbone is the most preferred method because it does not edit the video precisely.

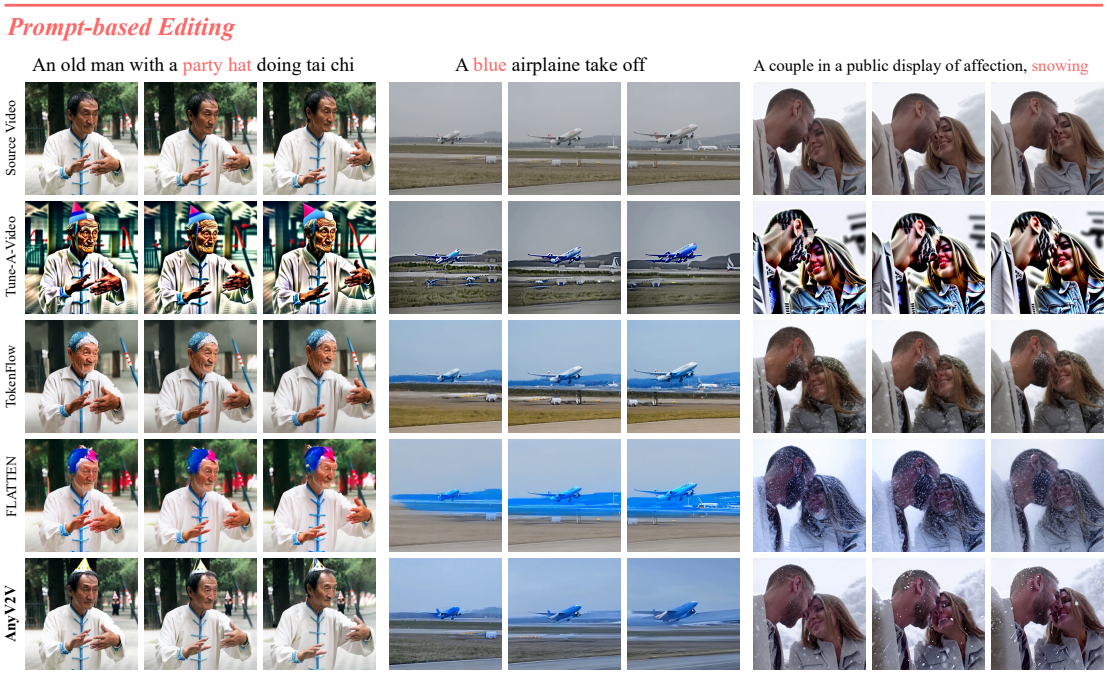


Figure 5.2: AnyV2V is robust in a wide range of prompt-based editing tasks while preserving the background. The results align the most with the text prompt and maintain high motion consistency.

Style Transfer, Subject-Driven Editing, and Identity Manipulation

Within the sphere of style transfer endeavors, the artistic techniques of Monet and Van Gogh are commonly examined and discussed. Traditional approaches to style transfer usually harness the model's built-in broad-spectrum knowledge, utilizing text descriptions to grasp and emulate these styles in video edits of high quality. While these conventional methods are effective for mimicking the styles of certain renowned artists, they often lack the precision needed to reflect the unique styles of other artists accurately. AnyV2V deviates from the standard practice of using only textual descriptions to guide style modifications. Leveraging the style transfer model NST [31], AnyV2V distinctly excels in precisely capturing the unique style of Vassily Kandinsky's work, a feat not accomplished by other methods, as shown in Figure 5.3. This degree of control provides artists with the novel capability to reference their own artworks for video editing purposes, thereby broadening the horizons for artistic creativity.

Table 5.2: Comparisons for three I2V models under AnyV2V framework on novel video editing tasks. Align: reference alignment; Overall: overall preference. **Bold**: best results; Underline: top-2.

Task	Reference-based Style Transfer		Subject-driven Editing		Identity Manipulation	
Image Editing Method	NST		AnyDoor		InstantID	
Image Editing Success Rate	≈90%		≈10%		≈80%	
Human Evaluation	Align ↑	Overall ↑	Align ↑	Overall ↑	Align ↑	Overall ↑
AnyV2V(SEINE)	92.3%	30.8%	48.4%	15.2%	72.7%	18.2%
AnyV2V(ConsistI2V)	38.4%	10.3%	63.6%	42.4%	72.7%	27.3%
AnyV2V(I2VGen-XL)	**100.0%**	**76.9%**	**93.9%**	**84.8%**	**90.1%**	**45.4%**

For Subject-Driven Editing, we utilized the subject-oriented image editing model AnyDoor for initial frame modifications. AnyDoor enables the substitution of any object within a target image with a subject from a singular reference image. Observations from Figure 5.3 indicate that AnyV2V is capable of creating videos with high motion consistency during subject-specific object replacement. For instance, in the first scenario, AnyV2V adeptly swaps a cat for a dog, mirroring the reference image while preserving the motion and background continuity seen in the original video. In another instance, it successfully exchanges a car with a preferred model, ensuring the edited video retains the original's rotational dynamics.

For identity manipulation, AnyV2V can be combined with InstantID [97] and ControlNet [109] to facilitate the alteration of a person's identity in the edited video, as exemplified in Figure 5.3. To our knowledge, this is the first instance of such a versatile application in video editing models. We report both human evaluation results and find that AnyV2V with I2VGen-XL backbone is the most preferred method in terms of reference alignment and overall performance, plotted in Table 5.1 and Table 5.2.

We find that AnyV2V(I2VGen-XL) tends to be the most robust both qualitatively and quantitatively. It has a good generalization ability to produce consistent motions in the video with high visual quality. AnyV2V(ConsistI2V) can generate consistent motion, but sometimes the watermark would appear due to its training data, thus harming the visual quality. AnyV2V(SEINE)'s generalization ability is relatively weaker but still produces consistent and high-quality video if the motion is simple enough, such as a person walking.

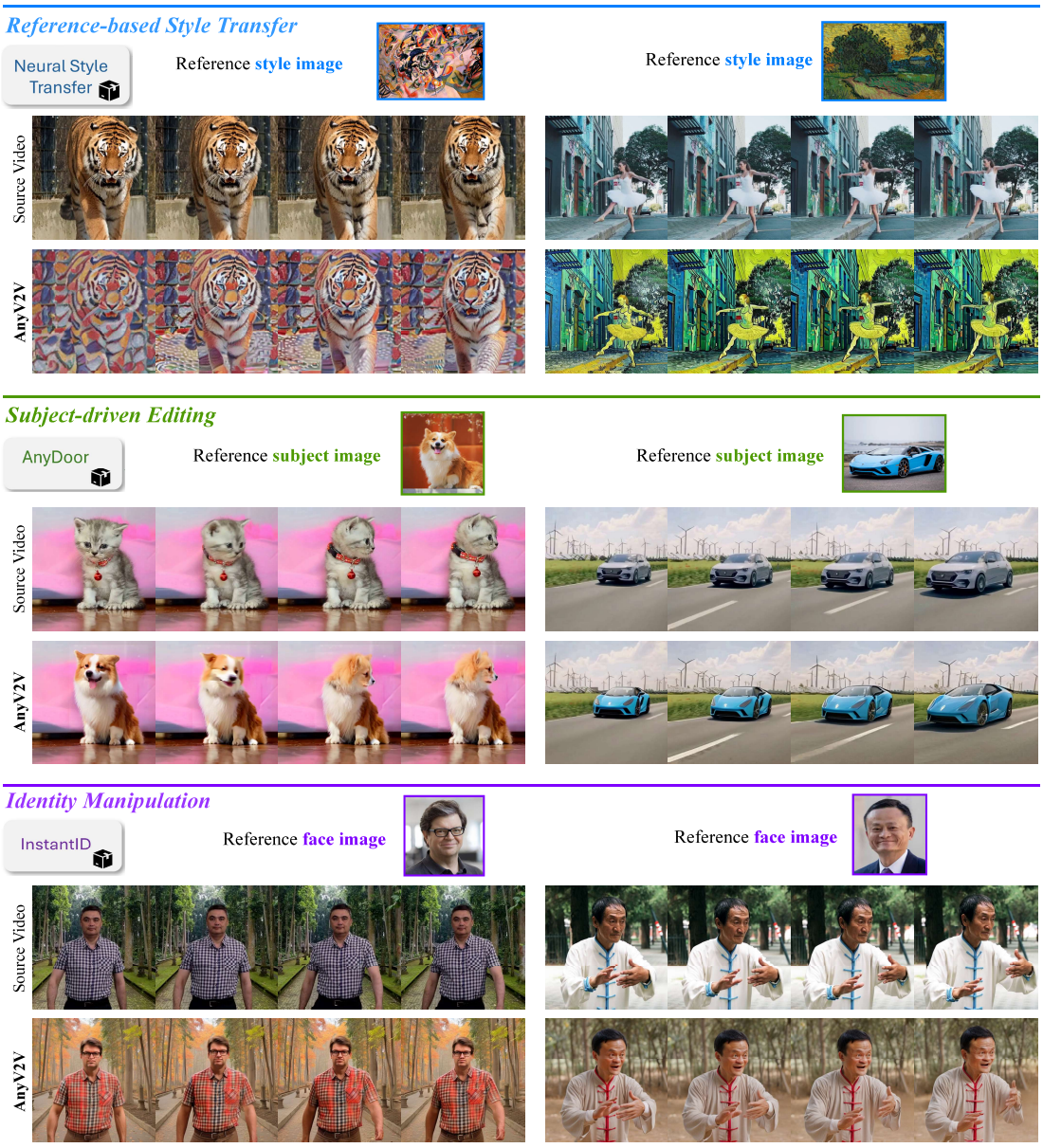


Figure 5.3: With different image editing models, AnyV2V can achieve a wide range of editing tasks, including reference-based style transfer, subject-driven editing, and identity manipulation.

5.2.2 Ablation Study

In this ablation study, we conduct experiment on AnyV2V with I2VGen-XL backbone.

Iteratively Disabling core components

To verify the effectiveness of our design choices, we conduct an ablation study by iteratively disabling the three core components in our model: temporal feature injection, spatial feature injection, and DDIM inverted latent as initial noise. We use a subset of 20 samples in this ablation study and report both the frame-wise consistency results using CLIP-Image score in Table 5.3 and qualitative comparisons in Figure 5.4.

Effectiveness of Temporal Feature Injection. As per the findings, when temporal feature injection is disabled in AnyV2V, there is a marginal increase in the CLIP-Image score value. However, the edited videos often exhibit reduced fidelity to the motion depicted in the source video. For instance, in the "couple sitting" scenario's second frame (found in the third row and second column of the right panel in Figure 5.4), the action of the woman lifting her leg in the source video is not replicated in the edited video when temporal injection is omitted. Conversely, even when the video style undergoes a complete transformation, AnyV2V with temporal injection effectively preserves subtle motions in the edited video.

Effectiveness of Spatial Feature Injection. As depicted in Table 5.3, a decline in the CLIP-Image score is evident when the spatial feature injection mechanisms are excluded from our model. This suggests that the edited videos lack smooth transitions across frames and exhibit more inconsistencies in appearance and motion. This is further exemplified in the third row of Figure 5.4, where the absence of spatial feature injection frequently leads to inaccuracies in subject appearance and pose (as seen in the "ballet dancing" case) and a degradation in background appearance (as observed in the "couple sitting" case). These findings underscore the importance of spatial feature injection mechanisms in enhancing editing outcomes, indicating that solely generating edited videos from the DDIM inverted noise is insufficient to fully preserve the structures of the source video.

DDIM Inverted Noise as Structural Guidance. A notable decline in CLIP-Image scores and a considerable degradation in visual quality are noted in both instances illustrated in Figure 5.4 when the initial DDIM inverted noise is substituted with random noise

Table 5.3: Ablation study results for AnyV2V. T. Injection and S. Injection correspond to temporal and spatial feature injection mechanisms, respectively.

Model	CLIP-Image ↑
AnyV2V	0.9648
AnyV2V w/o T. Injection	0.9652
AnyV2V w/o T. Injection & S. Injection	0.9637
AnyV2V w/o T. Injection & S. Injection & DDIM Inversion	0.9607

during sampling. This highlights that the I2V generation models exhibit reduced proficiency in animating the input image when faced with editing prompts that are entirely out-of-domain. It underscores the critical role of the DDIM inverted noise as the structural guide for the edited videos, emphasizing its significance in maintaining the integrity and quality of the editing process.



Figure 5.4: Visual comparisons after disabling temporal feature injection (T.I.), spatial feature injection (S.I.) and DDIM inverted initial noise (D.I.).

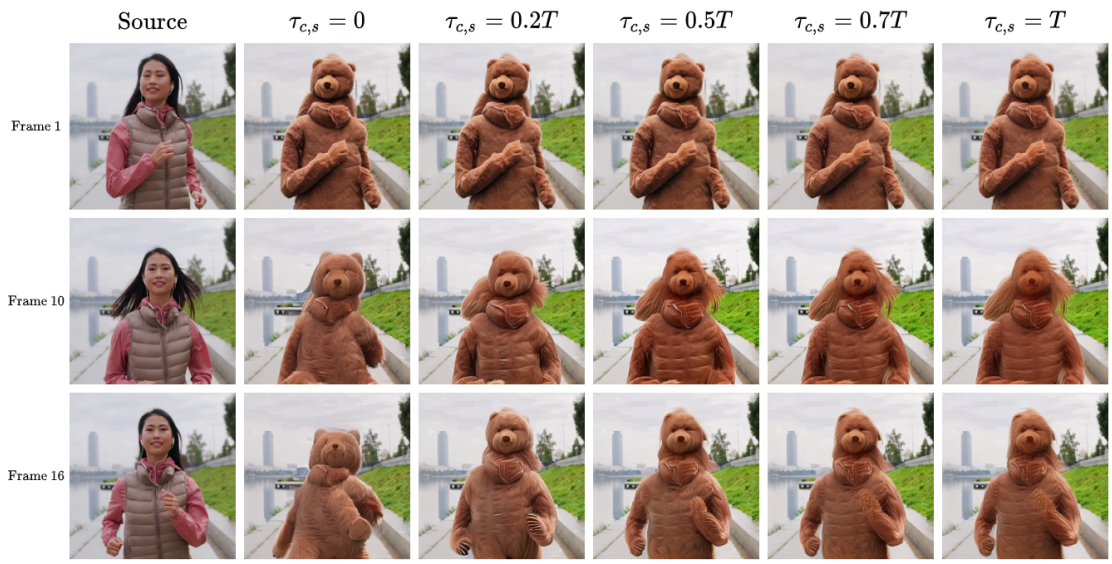


Figure 5.5: Hyperparameter study of Spatial Feature injection.

Hyperparameter Study on Spatial and Temporal feature injection

Effect of Spatial Injection Thresholds τ_{conv}, τ_{sa}. We examine the impact of deactivating spatial feature injection or adjusting the values of τ_{conv} and τ_{sa} in video editing, presenting the visual outcomes in Figure 5.5. Our observations reveal that without spatial feature injection, the edited videos deviate from the original video's layout and motion. Conversely, when the thresholds for spatial feature injection are excessively high, the edited videos become distorted by intricate details from the source video (such as textures from the woman's down jacket in Figure 5.5). By setting $\tau_{conv} = \tau_{sa} = 0.2T$, we achieve the desired editing results in our experiments.

Effect of Temporal Injection Threshold τ_{ta}. We investigate the impact of the hyperparameter τ_{ta} for temporal feature injection across various configurations, showcasing the outcomes in Figure 5.6. Our analysis reveals that when $\tau_{ta} < 0.5T$ (where T represents the total denoising steps), the motion guidance is insufficient, resulting in partially aligned motion with the source video, despite the motion being coherent and smooth. Conversely, with $\tau_{ta} > 0.5T$, the generated video better adheres to the motion but introduces distortion. Through our experiments, setting $\tau_{ta} = 0.5T$ strikes an optimal balance, ensuring alignment, consistency in motion, and fidelity in the video output.

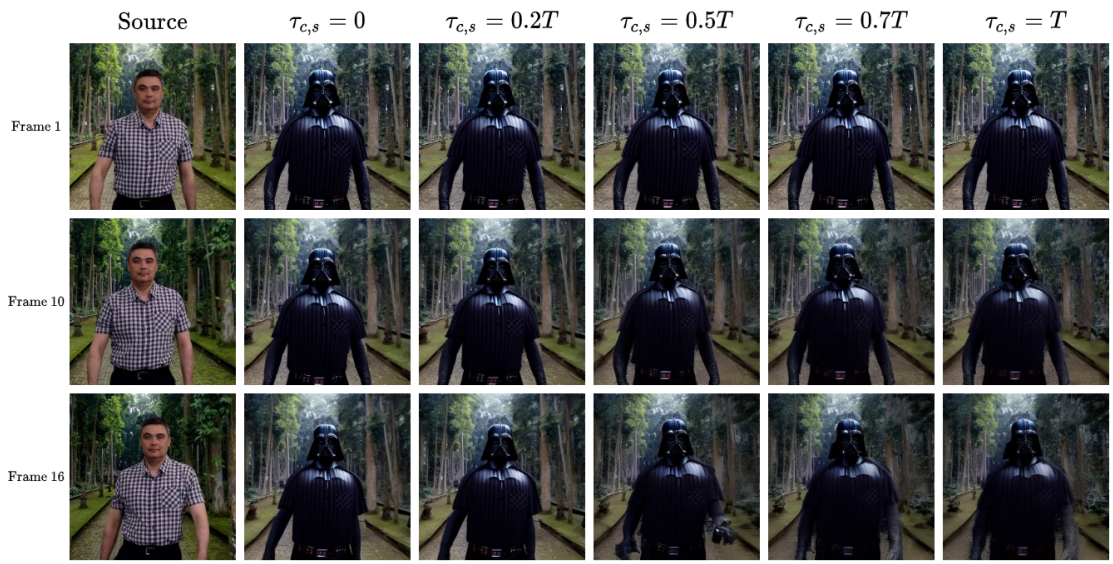


Figure 5.6: Hyperparameter study of Temporal Feature injection.

5.3 Limitation and Remedies

Long Video Editing. Most I2V models are trained on 16 frames, which means AnyV2V edits would be constrained to 16 frames edit. We explored the possible methods to allow AnyV2V to edit videos that is longer than 16 frames (2 seconds), without any extra training procedure. In the literature of FreeNoise ([78]), they leverage the modification of the initial latent to generate a video more than 4 times longer. It is tuning-free and can be adapted to most mainstream Text-To-Video and Image-To-Video diffusion models. In detail, it duplicates the initial latent into multiple chunks, then shuffles each chunk of duplicated latent. Finally, the initial latent and duplicated chunks are concatenated together, feeding into a T2V or I2V model as input. The trick was referred as Noise Rescheduling, and the motion feature injection allows video to generate much longer videos with improved frame consistency. This inspired us to directly compute the inversion on a lengthy video, forcing an Image-to-Video (I2V) model to infer long frames of output. Intuitively, this method would not work at all since the model is trained on 16 frames (2 seconds with 8 frames each), it is expected to yield poor results as the model should not be able to keep the consistency for more than 16 frames. However, the inverted latent should contain enough temporal and semantic information to allow the generated video to maintain the temporal and semantic consistency. Our experiment shows that this method allows video editing up to 16 seconds (128 frames) while maintaining the edit's visual quality and motion consistency.

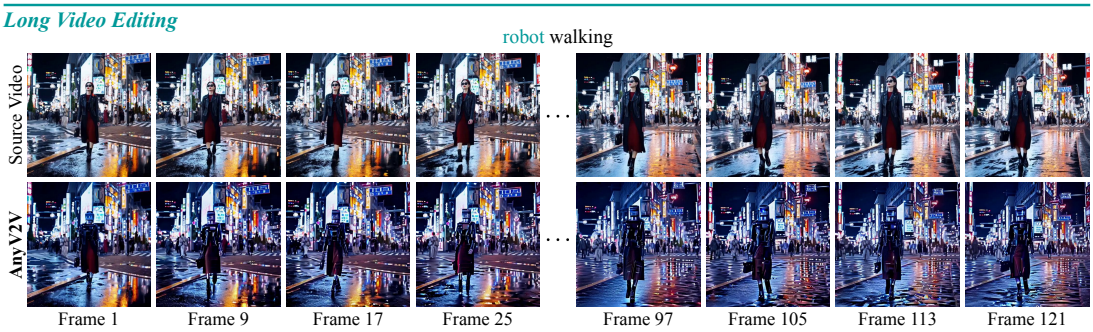


Figure 5.7: Performance of Inversion approach with AnyV2V on 128 frames edit.

Inaccurate Edit from Image Editing Models. Since our approach hinges on an initial frame adjustment, we utilize image editing models. Nevertheless, the existing cutting-edge models are not sufficiently developed to consistently execute precise edits [51]. For instance, during a task involving subject-driven video editing, we observed that AnyDoor [21] often necessitates multiple attempts to achieve a satisfactory editing outcome. This process involves the manual selection of a well-edited frame. We anticipate that forthcoming advancements in image editing models will streamline this process, reducing the required effort.

Limited ability of I2V models. Our research revealed that our technique struggles to replicate the motion of the source video when it is rapid (for example, billiard balls colliding at high speed) or intricate (such as a person cutting their hair). One potential explanation is that prevalent I2V models are typically trained on slow-motion footage, which limits their capacity to recreate swift or intricate movements even with motion guidance. We foresee that the development of a robust I2V model could resolve this challenge. One possible solution would be adapting our methods to Stable Video Diffusion [10] for stronger generalization ability.

5.4 Key Takeaway

The AnyV2V framework begins by processing a given source video. Initially, an image editing technique is applied to the first frame of the video, tailored to the specific editing task at hand. Such edit can be performed by a black-box image editing model or human

effort. Following this, the source video is converted back into initial noise latent through the inversion process with DDIM sampling. Then in the denoising process, both spatial features, self-attention features, and temporal attention features are injected into the I2V model decoder layers. The inverted latent is used as the initial latent during the denoising process and the edited first frame is fed into the I2V model as the control signal. Each features and attention extracted during the process are strategically integrated into the corresponding decoder layers of the model throughout this denoising stage. This produces high quality video edits with precise control.

The innovation part of the work is that this is the first work to propose such paradigm in aligning image editing and video editing problem with I2V models. The ability comes from not just the off-the-shelf image editing model but also the generalization power from the I2V model. This implies our proposed method is simple and very effective that we can achieve such high quality video editing by leveraging the research from the image editing problem, without any fine-tuning process on the video model. We showed that the results would improve when equipped with stronger Image editing models and I2V models. We believe AnyV2V serves as a good baseline and provides insights into the video editing task.

Chapter 6

Conclusion

This thesis conducted a thorough investigation of visual synthesis research, with a specific emphasis on two key domains: image generation/editing and video editing. By utilizing ImagenHub and VIEScore, we uncover both the capabilities and robustness on SOTA image synthesis model through standardized human evaluation, also explored the possible solution in image evaluation with llm-as-a-judge. Then with AnyV2V, we aligned the image editing and video editing problem and succeed in performing complex video editing tasks without necessitating significant retraining efforts.

ImagenHub served as a rigorous benchmark for evaluating common image synthesis tasks, which includes a evaluation dataset, an inference library, and rating guidelines to standardize the human rating process. Through extensive human evaluation, we observed that the existing models' performance is generally unsatisfying, especially in image editing. Complementarily, VIEScore leverages the Multimodal LLMs to evaluate synthesized images with instruction, enable the reasoning ability in image evaluation metrics. Our experiments showed the reliability of such metric increases when a stronger Multimodal LLM backbone is used, and it can perform on par with human ratings. This marked the possibility of employing explainable visual metrics in the future.

On the other hand, AnyV2V aligned image editing and video editing problem with I2V models. This approach not only provided a simple yet effective solution to edit video without fine-tuning but also allow precise control in the video edit. Our experiment showed that the results can be improved when plugged with a stronger I2V model backbone, and the control can be off-loaded to an image editing model. AnyV2V's success marked a novel paradigm in overcoming the lack of controls in general video editing.

In essence, we identified the true progress in the field of visual content generation and

manipulation, standardizing and improving its evaluation, and bridging image and video editing problem. This thesis proposed three paradigms for future explorations, aiming to enhance the robust control in visual generation and manipulation.

I hope the insights gained from these studies can illuminate the current state of visual generation and manipulation, highlighting both its strengths and areas needing improvement. With ongoing innovation and research, we expect the development of more advanced models that will overcome current limitations, further transforming our interaction with and comprehension of the capabilities of image synthesis.

References

[1] Stability AI. CosXL. https://huggingface.co/stabilityai/cosxl, 2024. Accessed on: 2024-04-13.

[2] Jie An, Songyang Zhang, Harry Yang, Sonal Gupta, Jia-Bin Huang, Jiebo Luo, and Xi Yin. Latent-shift: Latent diffusion with temporal shift for efficient text-to-video generation. *arXiv preprint arXiv:2304.08477*, 2023.

[3] Omri Avrahami, Dani Lischinski, and Ohad Fried. Blended diffusion for text-driven editing of natural images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18208–18218, 2022.

[4] Anas Awadalla, Irena Gao, Josh Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Gadre, Shiori Sagawa, Jenia Jitsev, Simon Kornblith, Pang Wei Koh, Gabriel Ilharco, Mitchell Wortsman, and Ludwig Schmidt. Openflamingo: An open-source framework for training large autoregressive vision-language models. *arXiv preprint arXiv:2308.01390*, 2023.

[5] Jianhong Bai, Tianyu He, Yuchi Wang, Junliang Guo, Haoji Hu, Zuozhu Liu, and Jiang Bian. Uniedit: A unified tuning-free framework for video motion and appearance editing. *arXiv preprint arXiv:2402.13185*, 2024.

[6] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 2023.

[7] Omer Bar-Tal, Hila Chefer, Omer Tov, Charles Herrmann, Roni Paiss, Shiran Zada, Ariel Ephrat, Junhwa Hur, Yuanzhen Li, Tomer Michaeli, et al. Lumiere: A space-time diffusion model for video generation. *arXiv preprint arXiv:2401.12945*, 2024.

[8] Omer Bar-Tal, Dolev Ofri-Amar, Rafail Fridman, Yoni Kasten, and Tali Dekel. Text2live: Text-driven layered image and video editing. In *European Conference on Computer Vision*, pages 707–723, 2022.

[9] Rohan Bavishi, Erich Elsen, Curtis Hawthorne, Maxwell Nye, Augustus Odena, Arushi Somani, and Sağnak Taşırlar. Introducing our multimodal models, 2023.

[10] A. Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, and Dominik Lorenz. Stable video diffusion: Scaling latent video diffusion models to large datasets. *ArXiv*, abs/2311.15127, 2023.

[11] Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22563–22575, 2023.

[12] Tim Brooks, Aleksander Holynski, and Alexei A. Efros. Instructpix2pix: Learning to follow image editing instructions. In *CVPR*, 2023.

[13] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, T. J. Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeff Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. *ArXiv*, abs/2005.14165, 2020.

[14] Mathilde Caron, Hugo Touvron, Ishan Misra, Herv'e J'egou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9630–9640, 2021.

[15] Duygu Ceylan, Chun-Hao Paul Huang, and Niloy J Mitra. Pix2video: Video editing using image diffusion. 2023.

[16] Haoxin Chen, Menghan Xia, Yingqing He, Yong Zhang, Xiaodong Cun, Shaoshu Yang, Jinbo Xing, Yaofang Liu, Qifeng Chen, Xintao Wang, et al. Videocrafter1: Open diffusion models for high-quality video generation. *arXiv preprint arXiv:2310.19512*, 2023.

[17] Haoxin Chen, Yong Zhang, Xiaodong Cun, Menghan Xia, Xintao Wang, Chao Weng, and Ying Shan. Videocrafter2: Overcoming data limitations for high-quality video diffusion models. *arXiv preprint arXiv:2401.09047*, 2024.

[18] Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Yue Wu, Zhongdao Wang, James T. Kwok, Ping Luo, Huchuan Lu, and Zhenguo Li. Pixart-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis. *ArXiv*, abs/2310.00426, 2023.

[19] Shoufa Chen, Mengmeng Xu, Jiawei Ren, Yuren Cong, Sen He, Yanping Xie, Animesh Sinha, Ping Luo, Tao Xiang, and Juan-Manuel Perez-Rua. Gentron: Delving deep into diffusion transformers for image and video generation. *arXiv preprint arXiv:2312.04557*, 2023.

[20] Wenhu Chen, Hexiang Hu, Yandong Li, Nataniel Rui, Xuhui Jia, Ming-Wei Chang, and William W Cohen. Subject-driven text-to-image generation via apprenticeship learning. *NeurIPS*, 2023.

[21] Xi Chen, Lianghua Huang, Yu Liu, Yujun Shen, Deli Zhao, and Hengshuang Zhao. Anydoor: Zero-shot object-level image customization. *arXiv preprint arXiv:2307.09481*, 2023.

[22] Xinyuan Chen, Yaohui Wang, Lingjun Zhang, Shaobin Zhuang, Xin Ma, Jiashuo Yu, Yali Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. Seine: Short-to-long video diffusion model for generative transition and prediction. *arXiv preprint arXiv:2310.20700*, 2023.

[23] Yuren Cong, Mengmeng Xu, Christian Simon, Shoufa Chen, Jiawei Ren, Yanping Xie, Juan-Manuel Perez-Rua, Bodo Rosenhahn, Tao Xiang, and Sen He. Flatten: optical flow-guided attention for consistent text-to-video editing. *arXiv preprint arXiv:2310.05922*, 2023.

[24] Guillaume Couairon, Jakob Verbeek, Holger Schwenk, and Matthieu Cord. Diffedit: Diffusion-based semantic image editing with mask guidance. In *The Eleventh International Conference on Learning Representations*, 2022.

[25] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Albert Li, Pascale Fung, and Steven C. H. Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. *ArXiv*, abs/2305.06500, 2023.

[26] deep floyd.ai. If by deepfloyd lab at stabilityai, 2023.

[27] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.

[28] Weixi Feng, Xuehai He, Tsu-Jui Fu, Varun Jampani, Arjun Reddy Akula, Pradyumna Narayana, Sugato Basu, Xin Eric Wang, and William Yang Wang. Training-free structured diffusion guidance for compositional text-to-image synthesis. In *The Eleventh International Conference on Learning Representations*, 2023.

[29] Joseph L Fleiss and Jacob Cohen. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and psychological measurement*, 33(3):613–619, 1973.

[30] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*, 2022.

[31] Leon A. Gatys, Alexander S. Ecker, and Matthias Bethge. A neural algorithm of artistic style. *CoRR*, abs/1508.06576, 2015.

[32] Songwei Ge, Seungjun Nah, Guilin Liu, Tyler Poon, Andrew Tao, Bryan Catanzaro, David Jacobs, Jia-Bin Huang, Ming-Yu Liu, and Yogesh Balaji. Preserve your own correlation: A noise prior for video diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22930–22941, 2023.

[33] Michal Geyer, Omer Bar-Tal, Shai Bagon, and Tali Dekel. Tokenflow: Consistent diffusion features for consistent video editing. *arXiv preprint arxiv:2307.10373*, 2023.

[34] Golnaz Ghiasi, Honglak Lee, Manjunath Kudlur, Vincent Dumoulin, and Jonathon Shlens. Exploring the structure of a real-time, arbitrary neural artistic stylization network. *CoRR*, abs/1705.06830, 2017.

[35] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks, 2014.

[36] Jing Gu, Yilin Wang, Nanxuan Zhao, Tsu-Jui Fu, Wei Xiong, Qing Liu, Zhifei Zhang, He Zhang, Jianming Zhang, HyunJoon Jung, et al. Photoswap: Personalized subject swapping in images. *arXiv preprint arXiv:2305.18286*, 2023.

[37] Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. In *The Twelfth International Conference on Learning Representations*, 2023.

[38] Agrim Gupta, Lijun Yu, Kihyuk Sohn, Xiuye Gu, Meera Hahn, Li Fei-Fei, Irfan Essa, Lu Jiang, and José Lezama. Photorealistic video generation with diffusion models. *arXiv preprint arXiv:2312.06662*, 2023.

[39] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7514–7528, 2021.

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

[41] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, et al. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022.

[42] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[43] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

[44] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *arXiv:2204.03458*, 2022.

[45] Berthold KP Horn and Brian G Schunck. Determining optical flow. *Artificial intelligence*, 17(1-3):185–203, 1981.

[46] Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2021.

[47] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. *CVPR*, 2017.

[48] Jin-Hwa Kim, Yunji Kim, Jiyoung Lee, Kang Min Yoo, and Sang-Woo Lee. Mutual information divergence: A unified metric for multimodal generative models. *Advances in Neural Information Processing Systems*, 35:35072–35086, 2022.

[49] Klaus Krippendorff. Computing krippendorff's alpha-reliability, 2011.

[50] Max Ku, Dongfu Jiang, Cong Wei, Xiang Yue, and Wenhu Chen. Viescore: Towards explainable metrics for conditional image synthesis evaluation. In *Proceedings of Annual Meeting of the Association for Computational Linguistics*, 2024.

[51] Max Ku, Tianle Li, Kai Zhang, Yujie Lu, Xingyu Fu, Wenwen Zhuang, and Wenhu Chen. Imagenhub: Standardizing the evaluation of conditional image generation models. In *The Twelfth International Conference on Learning Representations*, 2024.

[52] Max Ku, Cong Wei, Weiming Ren, Harry Yang, and Wenhu Chen. Anyv2v: A tuning-free framework for any video-to-video editing tasks. *arXiv preprint arXiv:2403.14468*, 2024.

[53] Wing-Fung Ku, Wan-Chi Siu, Xi Cheng, and H. Anthony Chan. Intelligent painter: Picture composition with resampling diffusion model, 2022.

[54] Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1931–1941, 2023.

[55] Tuomas Kynkäänniemi, Tero Karras, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Improved precision and recall metric for assessing generative models. *Advances in Neural Information Processing Systems*, 32, 2019.

[56] Tony Lee, Michihiro Yasunaga, Chenlin Meng, Yifan Mai, Joon Sung Park, Agrim Gupta, Yunzhi Zhang, Deepak Narayanan, Hannah Benita Teufel, Marco Bellagente, et al. Holistic evaluation of text-to-image models. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023.

[57] Dongxu Li, Junnan Li, and Steven CH Hoi. Blip-diffusion: Pre-trained subject representation for controllable text-to-image generation and editing. *arXiv preprint arXiv:2305.14720*, 2023.

[58] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*, 2022.

[59] Tianle Li, Max Ku, Cong Wei, and Wenhu Chen. Dreamedit: Subject-driven image editing. *Transactions on Machine Learning Research*, 2023.

[60] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *ArXiv*, abs/2304.08485, 2023.

[61] Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. What makes good in-context examples for gpt-3? In *Workshop on Knowledge Extraction and Integration for Deep Learning Architectures; Deep Learning Inside Out*, 2021.

[62] Yujie Lu, Xiujun Li, William Yang Wang, and Yejin Choi. Vim: Probing multimodal large language models for visual embedded instruction following. *ArXiv*, abs/2311.17647, 2023.

[63] Winfried Lötzsch, Max Reimann, Martin Büßemeyer, Amir Semmo, Jürgen Döllner, and Matthias Trapp. Wise: Whitebox image stylization by example-based learning. *ECCV*, 2022.

[64] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Guided image synthesis and editing with stochastic differential equations. In *International Conference on Learning Representations*, 2021.

[65] midjourney. Midjourney, 2023.

[66] Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models. *arXiv preprint arXiv:2211.09794*, 2022.

[67] Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6038–6047, 2023.

[68] Alexander Quinn Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob Mcgrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. In *International Conference on Machine Learning*, pages 16784–16804. PMLR, 2022.

[69] OpenAI. Video generation models as world simulators.

[70] OpenAI. Gpt-4 technical report, 2023.

[71] openai. Improving image generation with better captions, 2023.

[72] openjourney.ai. Openjourney is an open source stable diffusion fine tuned model on midjourney images, 2023.

[73] Dong Huk Park, Samaneh Azadi, Xihui Liu, Trevor Darrell, and Anna Rohrbach. Benchmark for compositional text-to-image synthesis. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*, 2021.

[74] Gaurav Parmar, Krishna Kumar Singh, Richard Zhang, Yijun Li, Jingwan Lu, and Jun-Yan Zhu. Zero-shot image-to-image translation. In *ACM SIGGRAPH 2023 Conference Proceedings*, pages 1–11, 2023.

[75] Archit Parnami and Minwoo Lee. Learning from few examples: A summary of approaches to few-shot learning. *ArXiv*, abs/2203.04291, 2022.

[76] Chenyang Qi, Xiaodong Cun, Yong Zhang, Chenyang Lei, Xintao Wang, Ying Shan, and Qifeng Chen. Fatezero: Fusing attentions for zero-shot text-based video editing. *arXiv:2303.09535*, 2023.

[77] Can Qin, Shu Zhang, Ning Yu, Yihao Feng, Xinyi Yang, Yingbo Zhou, Huan Wang, Juan Carlos Niebles, Caiming Xiong, Silvio Savarese, et al. Unicontrol: A unified diffusion model for controllable visual generation in the wild. *arXiv preprint arXiv:2305.11147*, 2023.

[78] Haonan Qiu, Menghan Xia, Yong Zhang, Yingqing He, Xintao Wang, Ying Shan, and Ziwei Liu. Freenoise: Tuning-free longer video diffusion via noise rescheduling, 2023.

[79] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[80] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 2021.

[81] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *ArXiv*, abs/2204.06125, 2022.

[82] Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy Lillicrap, Jean-baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024.

[83] Weiming Ren, Harry Yang, Ge Zhang, Cong Wei, Xinrun Du, Stephen Huang, and Wenhu Chen. Consisti2v: Enhancing visual consistency for image-to-video generation. *arXiv preprint arXiv:2402.04324*, 2024.

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

[85] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22500–22510, 2023.

[86] runwayml. Stable diffusion inpainting, 2023.

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

[88] Mehdi SM Sajjadi, Olivier Bachem, Mario Lucic, Olivier Bousquet, and Sylvain Gelly. Assessing generative models via precision and recall. *Advances in neural information processing systems*, 31, 2018.

[89] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. *Advances in neural information processing systems*, 29, 2016.

[90] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265. PMLR, 2015.

[91] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

[92] stability.ai. Stable diffusion xl, 2023.

[93] Roman Suvorov, Elizaveta Logacheva, Anton Mashikhin, Anastasia Remizova, Arsenii Ashukha, Aleksei Silvestrov, Naejin Kong, Harshith Goka, Kiwoong Park, and Victor Lempitsky. Resolution-robust large mask inpainting with fourier convolutions. *arXiv preprint arXiv:2109.07161*, 2021.

[94] Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. Plug-and-play diffusion features for text-driven image-to-image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1921–1930, June 2023.

[95] Patrick von Platen, Suraj Patil, Anton Lozhkov, Pedro Cuenca, Nathan Lambert, Kashif Rasul, Mishig Davaadorj, and Thomas Wolf. Diffusers: State-of-the-art diffusion models. https://github.com/huggingface/diffusers, 2022.

[96] Haofan Wang, Qixun Wang, Xu Bai, Zekui Qin, and Anthony Chen. Instantstyle: Free lunch towards style-preserving in text-to-image generation. *arXiv preprint arXiv:2404.02733*, 2024.

[97] Qixun Wang, Xu Bai, Haofan Wang, Zekui Qin, and Anthony Chen. Instantid: Zero-shot identity-preserving generation in seconds. *arXiv preprint arXiv:2401.07519*, 2024.

[98] Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, Jiazheng Xu, Bin Xu, Juanzi Li, Yuxiao Dong, Ming Ding, and Jie Tang. Cogvlm: Visual expert for pretrained language models. *ArXiv*, abs/2311.03079, 2023.

[99] Xiang Wang, Hangjie Yuan, Shiwei Zhang, Dayou Chen, Jiuniu Wang, Yingya Zhang, Yujun Shen, Deli Zhao, and Jingren Zhou. Videocomposer: Compositional video synthesis with motion controllability. *Advances in Neural Information Processing Systems*, 36, 2024.

[100] Yaohui Wang, Xinyuan Chen, Xin Ma, Shangchen Zhou, Ziqi Huang, Yi Wang, Ceyuan Yang, Yinan He, Jiashuo Yu, Peiqing Yang, et al. Lavie: High-quality video generation with cascaded latent diffusion models. *arXiv preprint arXiv:2309.15103*, 2023.

[101] Zijie J. Wang, Evan Montoya, David Munechika, Haoyang Yang, Benjamin Hoover, and Duen Horng Chau. DiffusionDB: A large-scale prompt gallery dataset for text-to-image generative models. *arXiv:2210.14896 [cs]*, 2022.

[102] Bichen Wu, Ching-Yao Chuang, Xiaoyan Wang, Yichen Jia, Kapil Krishnakumar, Tong Xiao, Feng Liang, Licheng Yu, and Peter Vajda. Fairy: Fast parallelized instruction-guided video-to-video synthesis. *arXiv preprint arXiv:2312.13834*, 2023.

[103] Chen Henry Wu and Fernando De la Torre. A latent space of stochastic diffusion models for zero-shot image editing and guidance. In *ICCV*, 2023.

[104] Jay Zhangjie Wu, Yixiao Ge, Xintao Wang, Stan Weixian Lei, Yuchao Gu, Yufei Shi, Wynne Hsu, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7623–7633, 2023.

[105] Jinbo Xing, Menghan Xia, Yong Zhang, Haoxin Chen, Xintao Wang, Tien-Tsin Wong, and Ying Shan. Dynamicrafter: Animating open-domain images with video diffusion priors. *arXiv preprint arXiv:2310.12190*, 2023.

[106] Shuai Yang, Yifan Zhou, Ziwei Liu, , and Chen Change Loy. Rerender a video: Zero-shot text-guided video-to-video translation. In *ACM SIGGRAPH Asia Conference Proceedings*, 2023.

[107] Yan Zeng, Guoqiang Wei, Jiani Zheng, Jiaxin Zou, Yang Wei, Yuchen Zhang, and Hang Li. Make pixels dance: High-dynamic video generation. *arXiv preprint arXiv:2311.10982*, 2023.

[108] Kai Zhang, Lingbo Mo, Wenhu Chen, Huan Sun, and Yu Su. Magicbrush: A manually annotated dataset for instruction-guided image editing. *NeurIPS dataset and benchmark track*, 2023.

[109] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023.

[110] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018.

[111] Shiwei Zhang, Jiayu Wang, Yingya Zhang, Kang Zhao, Hangjie Yuan, Zhiwu Qin, Xiang Wang, Deli Zhao, and Jingren Zhou. I2vgen-xl: High-quality image-to-video synthesis via cascaded diffusion models. *arXiv preprint arXiv:2311.04145*, 2023.

[112] Daquan Zhou, Weimin Wang, Hanshu Yan, Weiwei Lv, Yizhe Zhu, and Jiashi Feng. Magicvideo: Efficient video generation with latent diffusion models. *arXiv preprint arXiv:2211.11018*, 2022.

Appendix A

Supplementary material of ImagenHub

This section contains supplementary figures and tables that would take up too much space in the main thesis content.

A.1 Visualization Tools from ImagenHub

We are publicly sharing the benchmarking results to display the true performance of each model, known as Imagen Museum. This transparency allows viewers to gain insights from the actual outcomes. Our framework also encourages future researchers to release their human evaluation set generation results publicly. By standardizing the human evaluation protocol, we aim to enhance the credibility of human evaluation results through greater experimental transparency.

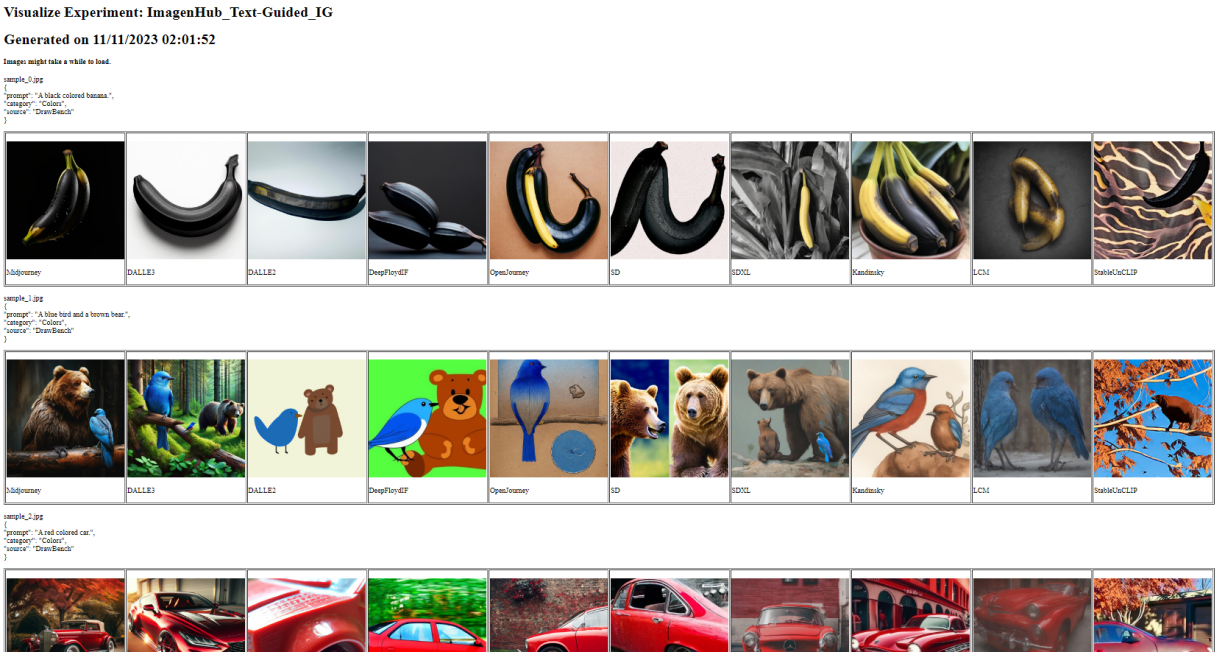


Figure A.1: The webpage design of Imagen Museum (Text-Guided Image Generation page).

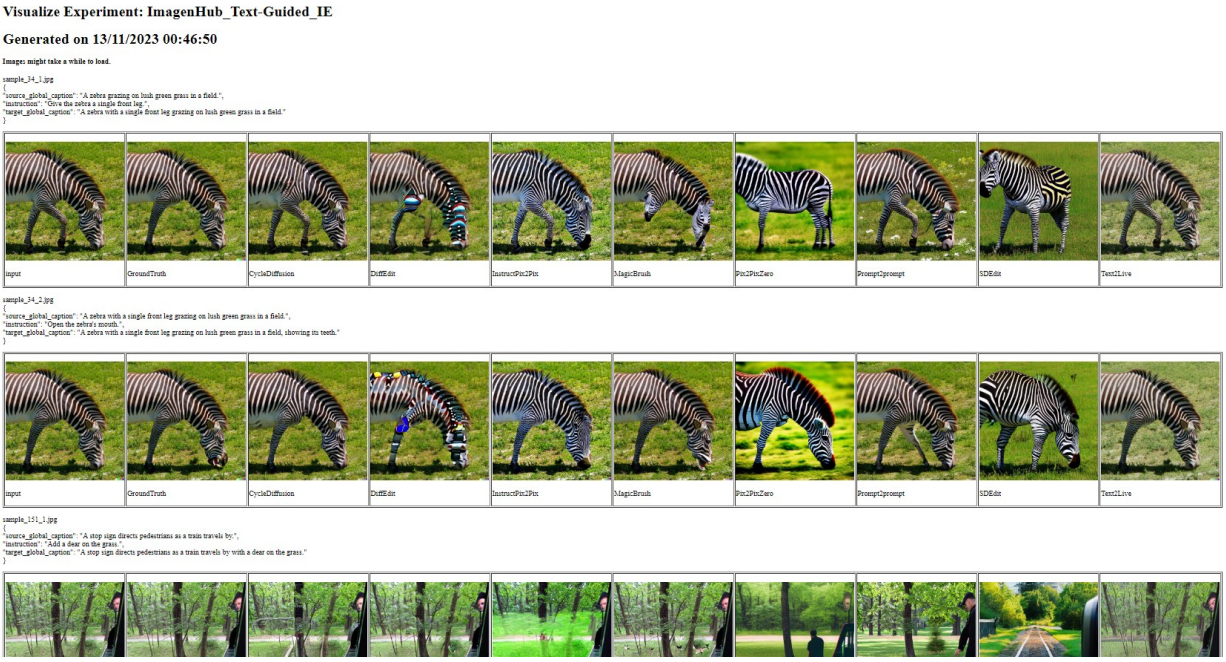


Figure A.2: The webpage design of Imagen Museum (Text-Guided Image Editing page).

Appendix B

Supplementary material of VIEScore

This section contains supplementary figures and tables that would take up too much space in the main thesis content.

B.1 Prompt Templates

The prompt from VIEScore is divided into two segments: the 'context prompt' and the 'rating prompt'. The ultimate prompt provided to the model is a combination of these two segments.

Context

You are a professional digital artist. You will have to evaluate the effectiveness of the AI-generated image(s) based on the given rules. You will have to give your output in this way (Keep your reasoning concise and short.):
{
"score" : [...],
"reasoning" : "..."
}

PQ Rating Prompt Template (for all tasks)

RULES:
The image is an AI-generated image. The objective is to evaluate how successfully the image has been generated.
On a scale 0 to 10:
A score from 0 to 10 will be given based on image naturalness.
(0 indicates that the scene in the image does not look natural at all or gives an unnatural feeling such as a wrong sense of distance, wrong shadow, or wrong lighting. 10 indicates that the image looks natural.)
A second score from 0 to 10 will rate the image artifacts.
(0 indicates that the image contains a large portion of distortion, watermarks, scratches, blurred faces, unusual body parts, or subjects not harmonized. 10 indicates the image has no artifacts.)
Put the score in a list such that output score = [naturalness, artifacts]

SC Rating Prompt Template (Text-Guided Image Generation)

RULES:
The image is an AI-generated image according to the text prompt. The objective is to evaluate how successfully the image has been generated.
On a scale 0 to 10:
A score from 0 to 10 will be given based on the success in following the prompt. (0 indicates that the AI-generated image does not follow the prompt at all. 10 indicates the AI-generated image follows the prompt perfectly.)
Put the score in a list such that output score = [score].
Text Prompt: ¡prompt¿

SC Rating Prompt Template (Text/Mask-Guided Image Editing)

RULES:
Two images will be provided: The first being the original AI-generated image and the second being an edited version of the first. The objective is to evaluate how successfully the editing instruction has been executed in the second image. Note that sometimes the two images might look identical due to the failure of the image edit.
On scale of 0 to 10:
A score from 0 to 10 will be given based on the success of the editing. (0 indicates that the scene in the edited image does not follow the editing instructions at all. 10 indicates that the scene in the edited image follows the editing instruction text perfectly.)
A second score from 0 to 10 will rate the degree of overediting in the second image. (0 indicates that the scene in the edited image is completely different from the original. 10 indicates that the edited image can be recognized as a minimally edited yet effective version of the original.)
Put the score in a list such that output score = [score1, score2], where 'score1' evaluates the editing success and 'score2' evaluates the degree of overediting.
Editing instruction: ¡instruction¿

SC Rating Prompt Template (Control-Guided Image Generation)

RULES:
Two images will be provided: The first being a processed image (e.g. Canny edges, openpose, grayscale, etc.) and the second being an AI-generated image using the first image as guidance. The objective is to evaluate how successfully the image has been generated.
On scale 0 to 10:
A score from 0 to 10 will be given based on the success in following the prompt. (0 indicates that the second image does not follow the prompt at all. 10 indicates the second image follows the prompt perfectly.)
A second score from 0 to 10 will rate how well the generated image is following the guidance image. (0 indicates that the second image does not follow the guidance at all. 10 indicates that the second image is following the guidance image.)
Put the score in a list such that output score = [score1, score2], where 'score1' evaluates the prompt and 'score2' evaluates the guidance.
Text Prompt: ¡prompt¿

SC Rating Prompt Template (Subject-Driven Image Generation)

RULES:
Two images will be provided: The first is a token subject image and the second is an AI-generated image using the first image as guidance. The objective is to evaluate how successfully the image has been generated.
On a scale of 0 to 10:
A score from 0 to 10 will be given based on the success in following the prompt. (0 indicates that the second image does not follow the prompt at all. 10 indicates the second image follows the prompt perfectly.)
A second score from 0 to 10 will rate how well the subject in the generated image resembles the token subject in the first image. (0 indicates that the subject in the second image does not look like the token subject at all. 10 indicates the subject in the second image looks exactly like the token subject.)
Put the score in a list such that output score = [score1, score2], where 'score1' evaluates the prompt and 'score2' evaluates the resemblance.
Text Prompt: ¡prompt¿

SC Rating Prompt Template (Subject-Guided Image Editing)

RULES:
Three images will be provided: The first image is an input image to be edited. The second image is a token subject image. The third image is an AI-edited image from the first image. it should contain a subject that looks like the subject in the second image. The objective is to evaluate how successfully the image has been edited.
On a scale 0 to 10:
A score from 0 to 10 will rate how well the subject in the generated image resembles the token subject in the second image. (0 indicates that the subject in the third image does not look like the token subject at all. 10 indicates the subject in the third image looks exactly like the token subject.)
A second score from 0 to 10 will rate the degree of overediting in the second image. (0 indicates that the scene in the edited image is completely different from the first image. 10 indicates that the edited image can be recognized as a minimally edited yet effective version of the original.)
Put the score in a list such that output score = [score1, score2], where 'score1' evaluates the resemblance and 'score2' evaluates the degree of overediting.
Subject: ¡subject¿

B.2 Correlations

Correlation to Human ratings Across All 7 Tasks			
	SC_{corr}	PQ_{corr}	$Overall_{corr}$
Human Raters	0.4700	0.4124	0.4558
VIEScore Backbones			
$GPT\text{-}4o_{0shot}$	0.4459	0.3399	0.4041
$GPT\text{-}4o_{1shot}$	0.4309	0.1167	0.3770
$Gemini\text{-}Pro_{0shot}$	0.3322	0.2675	0.3048
$Gemini\text{-}Pro_{1shot}$	0.3094	0.3070	0.3005
$GPT\text{-}4v_{0shot}$	0.3655	0.3092	0.3266
$GPT\text{-}4v_{1shot}$	0.2689	0.2338	0.2604
$LLaVA_{0shot}$	0.1046	0.0319	0.0925
$LLaVA_{1shot}$	0.1012	0.0138	0.0695
$Qwen\text{-}VL_{0shot}$	0.0679	0.0165	0.0920
$BLIP2_{0shot}$	0.0504	-0.0108	0.0622
$InstructBLIP_{0shot}$	0.0246	0.0095	0.0005
$Fuyu_{0shot}$	-0.0110	-0.0172	0.0154
$CogVLM_{0shot}$	-0.0228	0.0514	-0.0050
$OpenFlamingo_{0shot}$	-0.0037	-0.0102	-0.0122

Table B.1: Correlations across all tasks with different backbone models. We highlight the highest correlation numbers in green.

	SC_{corr}	PQ_{corr}	$Overall_{corr}$
Text-guided Image Generation (5 models)			
Human Raters	0.5044	0.3640	0.4652
CLIP-Score	-0.0817	-0.0114	-0.0881
VIEScore			
GPT-4o$_{0shot}$	0.4989	0.2495	0.3928
GPT-4o$_{1shot}$	0.5124	0.0336	0.4042
Gemini-Pro$_{0shot}$	0.5123	0.1842	0.4356
Gemini-Pro$_{1shot}$	0.4757	0.2206	0.4326
GPT-4v$_{0shot}$	0.4885	0.2379	0.4614
GPT-4v$_{1shot}$	0.4531	0.1770	0.3801
LLaVA$_{0shot}$	0.1809	0.0306	0.1410
LLaVA$_{1shot}$	0.1789	-0.0020	0.1309
Mask-guided Image Editing (4 models)			
Human Raters	0.5390	0.5030	0.4981
LPIPS	-0.1012	0.0646	-0.0694
VIEScore			
GPT-4o$_{0shot}$	0.5421	0.3469	0.4769
GPT-4o$_{1shot}$	0.5246	0.1272	0.4432
Gemini-Pro$_{0shot}$	0.4304	0.2839	0.3593
Gemini-Pro$_{1shot}$	0.4595	0.3170	0.4017
GPT-4v$_{0shot}$	0.4508	0.2859	0.4069
GPT-4v$_{1shot}$	0.4088	0.2352	0.3810
LLaVA$_{0shot}$	0.1180	-0.0531	0.0675
LLaVA$_{1shot}$	0.1263	-0.0145	0.1040
Text-guided Image Editing (8 models)			
Human Raters	0.4230	0.5052	0.4184
LPIPS	0.0956	0.2504	0.1142
VIEScore			
GPT-4o$_{0shot}$	0.4062	0.4863	0.3821
GPT-4o$_{1shot}$	0.3684	0.1939	0.3438
Gemini-Pro$_{0shot}$	0.2836	0.4291	0.2728
Gemini-Pro$_{1shot}$	0.2805	0.4657	0.2648
GPT-4v$_{0shot}$	0.2610	0.4274	0.2456
GPT-4v$_{1shot}$	0.2428	0.3402	0.2279
LLaVA$_{0shot}$	0.0448	0.0583	0.0273
LLaVA$_{1shot}$	0.0185	-0.0107	0.0258

Table B.2: Correlations to human ratings in different available methods. (Part 1)

	SC_{corr}	PQ_{corr}	$Overall_{corr}$
Subject-driven Image Generation (4 models)			
Human Raters	0.4780	0.3565	0.4653
DINO	0.4160	0.1206	0.4246
CLIP-I	0.2961	0.1694	0.3058
VIEScore			
GPT-4o$_{0shot}$	0.4806	0.2576	0.4637
GPT-4o$_{1shot}$	0.4685	-0.0171	0.4292
Gemini-Pro$_{0shot}$	0.2906	0.1765	0.2851
Gemini-Pro$_{1shot}$	0.3486	0.2800	0.3342
GPT-4v$_{0shot}$	0.3979	0.1903	0.3738
GPT-4v$_{1shot}$	0.2757	0.2261	0.2753
LLaVA$_{0shot}$	0.0326	-0.0303	0.1219
LLaVA$_{1shot}$	0.1334	0.0858	0.1248
Subject-driven Image Editing (3 models)			
Human Raters	0.4887	0.2986	0.4747
DINO	0.3022	-0.0381	0.3005
CLIP-I	0.2834	0.1248	0.2813
VIEScore			
GPT-4o$_{0shot}$	0.4800	0.3734	0.3268
GPT-4o$_{1shot}$	0.3862	0.1273	0.2797
Gemini-Pro$_{0shot}$	0.2187	0.3148	0.2234
Gemini-Pro$_{1shot}$	-0.0083	0.3181	0.0004
GPT-4v$_{0shot}$	0.3274	0.2960	0.1507
GPT-4v$_{1shot}$	-0.0255	0.1572	-0.0139
LLaVA$_{0shot}$	0.0360	-0.0073	0.0168
LLaVA$_{1shot}$	0.0587	-0.0249	0.0309

Table B.3: Correlations to human ratings in different available methods. (Part 2)

	SC_{corr}	PQ_{corr}	$Overall_{corr}$
Multi-concept Image Composition (3 models)			
Human Raters	0.5927	0.5145	0.5919
DINO	0.0979	-0.1643	0.0958
CLIP-I	0.1512	-0.0963	0.1498
VIEScore			
GPT-4o$_{0shot}$	0.4516	0.2751	0.4136
GPT-4o$_{1shot}$	0.4120	-0.0141	0.3523
Gemini-Pro$_{0shot}$	0.3557	0.1948	0.3314
Gemini-Pro$_{1shot}$	0.4151	0.1798	0.4131
GPT-4v$_{0shot}$	0.3209	0.3025	0.3346
GPT-4v$_{1shot}$	0.1859	0.1185	0.1918
LLaVA$_{0shot}$	0.1022	0.1194	0.1070
LLaVA$_{1shot}$	0.0828	0.0379	0.0293
Control-guided Image Generation (2 models)			
Human Raters	0.5443	0.5279	0.5307
LPIPS	0.3699	0.4204	0.4133
VIEScore			
GPT-4o$_{0shot}$	0.4972	0.4892	0.5439
GPT-4o$_{1shot}$	0.5544	0.3699	0.5238
Gemini-Pro$_{0shot}$	0.3254	0.3359	0.2960
Gemini-Pro$_{1shot}$	0.2677	0.4392	0.3240
GPT-4v$_{0shot}$	0.4360	0.4975	0.3999
GPT-4v$_{1shot}$	0.3892	0.4132	0.4237
LLaVA$_{0shot}$	0.2207	0.1060	0.1679
LLaVA$_{1shot}$	0.1121	0.0247	0.0416

Table B.4: Correlations to human ratings in different available methods. (Part 3)

Appendix C

Supplementary material of AnyV2V

C.1 Preliminary

Image-to-Video (I2V) Generation Models.

In this work, we focus on leveraging latent diffusion-based [84] I2V generation models for video editing. Given an input first frame I_1, a text prompt $\mathbf{s}$ and a noisy video latent $\mathbf{z}_t$ at time step t, I2V generation models recover a less noisy latent $\mathbf{z}_{t-1}$ using a denoising model $\epsilon_\theta(\mathbf{z}_t, I_1, \mathbf{s}, t)$ conditioned on both I_1 and $\mathbf{s}$. The denoising model ϵ_θ contains a set of spatial and temporal self-attention layers, where the self-attention operation can be formulated as:

$$Q = W^Q z, K = W^K z, V = W^V z, \tag{C.1}$$

$$\text{Attention}(Q, K, V) = \text{Softmax}(\frac{QK^\top}{\sqrt{d}})V, \tag{C.2}$$

where z is the input hidden state to the self-attention layer and W^Q, W^K and W^V are learnable projection matrices that map z onto query, key and value vectors, respectively. For spatial self-attention, z represents a sequence of spatial tokens from each frame. For temporal self-attention, z is composed of tokens located at the same spatial position across all frames.

DDIM Inversion.

The denoising process for I2V generation models from $\mathbf{z}_t$ to $\mathbf{z}_{t-1}$ can be achieved using the DDIM [91] sampling algorithm. The reverse process of DDIM sampling, known as DDIM inversion [67, 27], allows obtaining $\mathbf{z}_{t+1}$ from $\mathbf{z}_t$ such that $\mathbf{z}_{t+1} = \sqrt{\frac{\alpha_{t+1}}{\alpha_t}}\mathbf{z}_t + (\sqrt{\frac{1}{\alpha_{t+1}} - 1} - \sqrt{\frac{1}{\alpha_t} - 1}) \cdot \epsilon_\theta(\mathbf{z}_t, x_0, \mathbf{s}, t)$, where α_t is derived from the variance schedule of the diffusion process.

Plug-and-Play (PnP) Diffusion Features.

Tumanyan et al. [94] proposed PnP diffusion features for image editing, based on the observation that intermediate convolution features f and self-attention scores $A = \text{Softmax}(\frac{QK^\top}{\sqrt{d}})$ in a text-to-image (T2I) denoising U-Net capture the semantic regions (e.g. legs or torso of a human body) during the image generation process.

Given an input source image I^S and a target prompt P, PnP first performs DDIM inversion to obtain the image's corresponding noise $\{\mathbf{z}_t^S\}_{t=1}^T$ at each time step t. It then collects the convolution features $\{f_t^l\}$ and attention scores $\{A_t^l\}$ from some predefined layers l at each time step t of the backward diffusion process $\mathbf{z}_{t-1}^S = \epsilon_\theta(\mathbf{z}_t^S, \varnothing, t)$, where $\varnothing$ denotes the null text prompt during denoising.

To generate the edited image I^*, PnP starts from the initial noise of the source image (i.e. $\mathbf{z}_T^* = \mathbf{z}_T^S$) and performs feature injection during denoising: $\mathbf{z}_{t-1}^* = \epsilon_\theta(\mathbf{z}_t^*, P, t, \{f_t^l, A_t^l\})$, where $\epsilon_\theta(\cdot, \cdot, \cdot, \{f_t^l, A_t^l\})$ represents the operation of replacing the intermediate feature and attention scores $\{f_t^{l*}, A_t^{l*}\}$ with $\{f_t^l, A_t^l\}$. This feature injection mechanism ensures I^* to preserve the layout and structure from I^S while reflecting the description in P. To control the feature injection strength, PnP also employs two thresholds τ_f and τ_A such that the feature and attention scores are only injected in the first τ_f and τ_A denoising steps. Our method extends this feature injection mechanism to I2V generation models, where we inject features in convolution, spatial, and temporal attention layers.