

Lecture 7:

Condition-based Diffusion Models for MM Generation

Papers for Lecture 7 (Diffusion Models for MM Generation)

- P7-1: Diffusion Models for Image Generation: Presenter: Xing Naili; Reader: Chai Zenghao**
(Must-Read) J Ho, A Jain & P Abbeel. Denoising Diffusion Probabilistic Models. NeurIPS 2020
(Must-Read) J Song, C Meng & S Ermon. Denoising Diffusion Implicit Models. ICLR 2021.
(To-Read) Y Song, et al. Score-Based Generative Modeling through Stochastic Differential Equations. ICLR 2021.
- P7-2: Condition-based Diffusion Models: Presenter: Chen Xihao; Reader: Nguyen Thong Thanh**
(Must-Read) X Shen, et al. Fine Tuning Text-to-Image Diffusion Models for Fairness. ICLR 2024.
(Must-Read) R Rombach, et al. High-Res Image Synthesis with Latent Diffusion Models. CVPR 2022.
(To-Read, Best Paper) L Zhang, et al. Adding Conditional Control to Text-to-Image Diffusion Models. ICCV 2023.
- P7-3: Image/Video Editing & Personalization: Presenter: Lin Xinyu; Reader: Zheng Jingnan**
(Must-Read) A Hertz, et al. Prompt-to-Prompt Image Editing with Cross Attention Control. ICLR 2023.
(To-Read) H Ouyang, et al. CoDeF: Content Deformation Fields for Temporally Consistent Video Processing. arXiv 2023.
(Must-Read) N Ruiz, et al. DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation. CVPR 2023.

Text-Guided Image Generation

Definition:

Text-to-image generation is aimed at generating photo-realistic images according to the text description and has great potential for applications e.g., image editing, computer-aided design, entertainment interaction.

Applications:

Image
Generation

The bird is **short** and **stubby** with
yellow on its body

Dec



StackGAN (ICCV2017)

Image
Editing



A bird with **black eye rings** and **a black bill**,
with a **red crown** and a **red belly**.

Enc

Dec



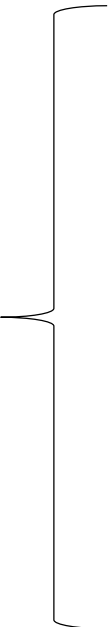
ManiGAN (CVPR2020)

Conditioned Visual Generation

- We have seen the focus on conditioned generations. What are the conditions?
 - Simple constraints: numeric, positional
 - Esthetic/ Artistic constraints
 - Symmetry constraints
 - Etc.

■ Classification of Visual Editing

- Essential for solving the last mile problem in visual generation
- Requires local changes/editing: partially supported by Dall-E-3:

| | | | | |
|--|------------------|---|------------------|-----------------|
|  | Local | object relations ✓ | color actions ✓ | counting text ✓ |
| | Global | changing background ✓ | style transfer ✓ | tone transfer ✓ |
| | Visual | reference (canny edge, sketch, exemplar) attributes | | |
| | Viewpoint | camera lens movement | | |

■ Local

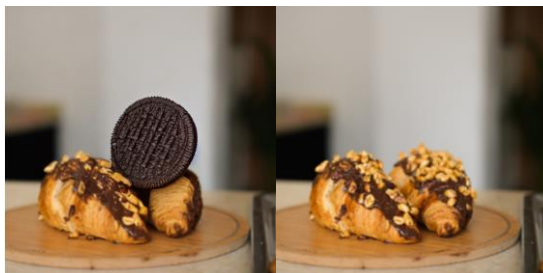
object alter

cat to dog



object delete

remove the cookie



object add

add "Car"



counting

only one gift in the ground



color alter

red to pink



relation

Put the gloves from the **right** of the fabric basket to the **left** of it



action

let the dog sit



texture modification

text: "English" to "Chinese"



■ Global

changing background

“Table” to “Galaxy”



tone transfer

“Fall” to “Winter”



style transfer

painting to “Children’s drawing”



stylize “Blanket”



■ Visual

reference



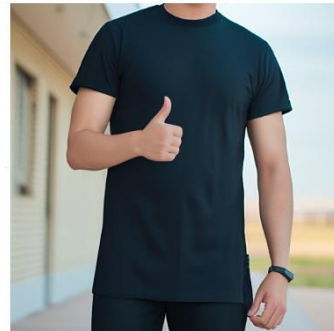
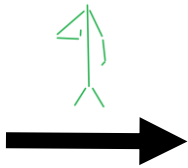
stroke



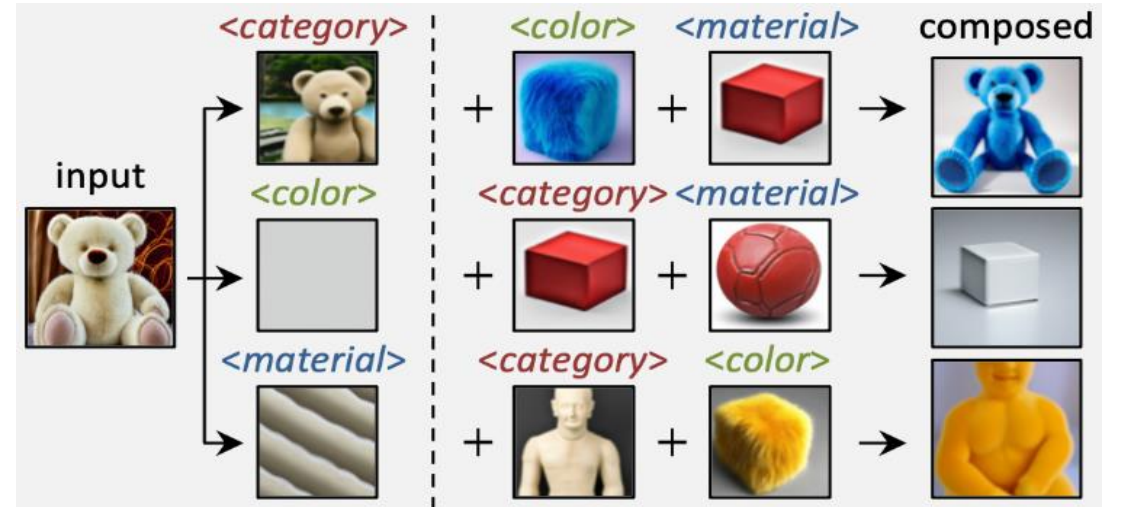
exemplar



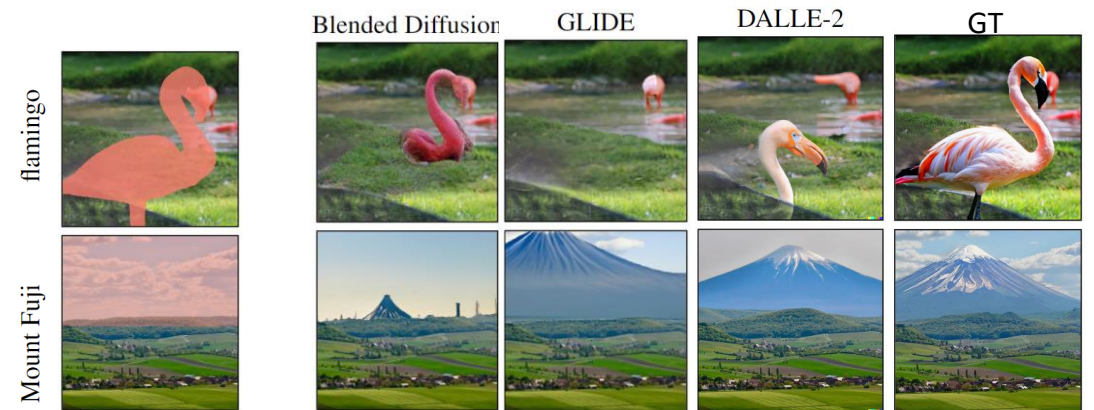
canny
edge



attributes



strict bounding



■ Implicit

make it more healthy



let's see it graduating



■ Viewpoint

Look further away



■ Interleaved Editing Instruction

Photo of Bill Gates with the same hand gesture as in the given image



, with a dog looks like this one in the image



1. **Implicit:** change the man to Bill Gates
2. **local:** add a dog look like this one in the image



Over-edit

MagicBrush504170 The background should be of a mountain.

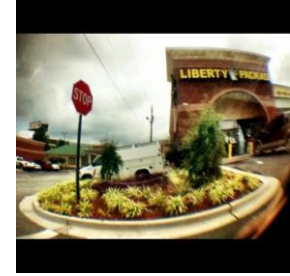


seed dataset is not clean

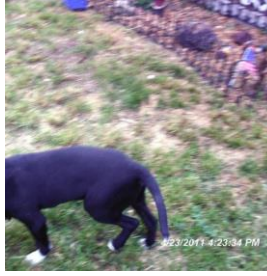
MagicBrush553990 Make the horse's coat dark black.



278535



483587



watermark

Remove DALLE 2 Watermark



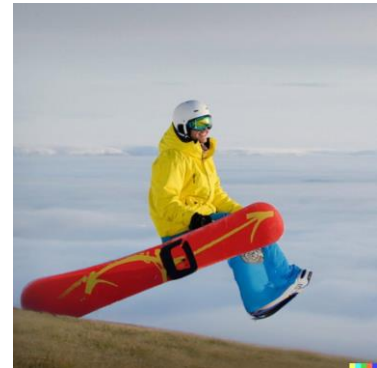
MagicBrush173530 What if it was another car on it?



MagicBrush413182 Have both kids be wearing red baseball caps



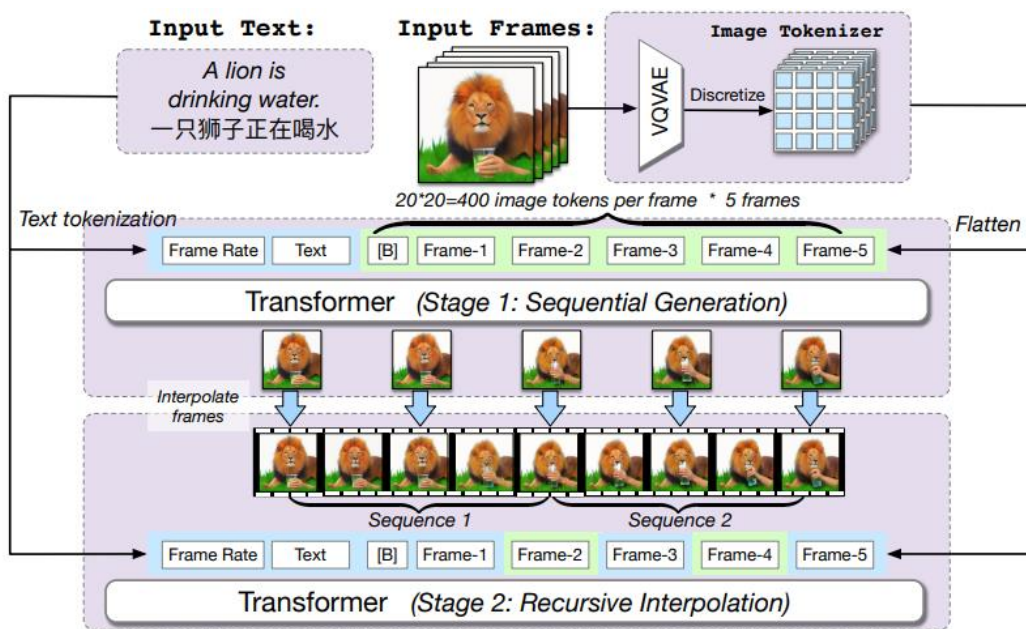
MagicBrush50058 let the snow pile be a grassy hill



Just random search those in the [first 100 pairs!](#)
And not showcase the all not good pairs

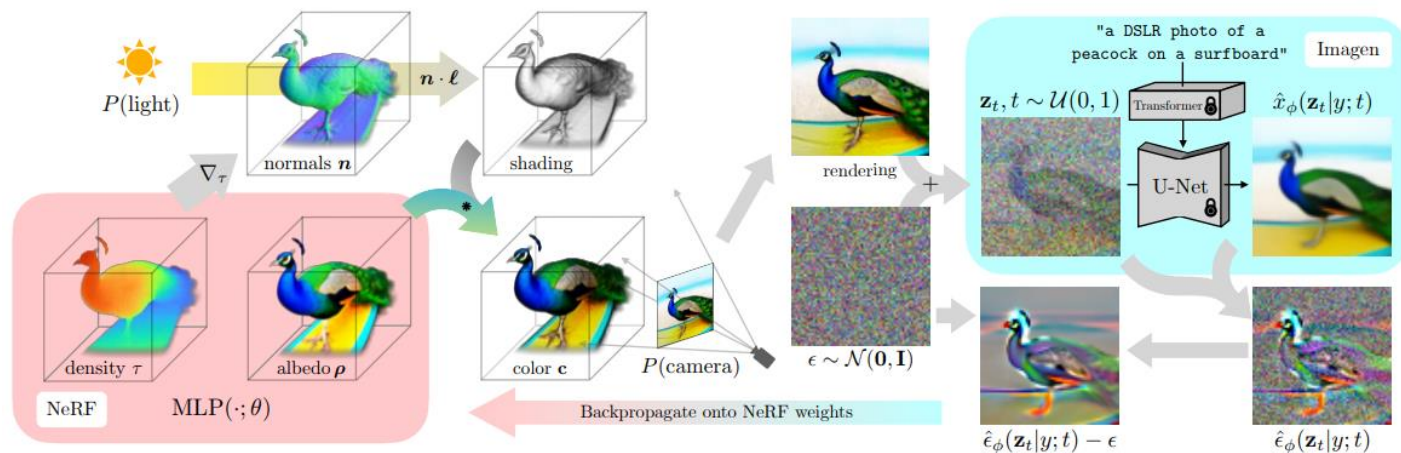
Text-Guided Multimodal Generation

Text-video Generation:



CogVideo

Text-3D Generation:



Dreamfusion

The Emergence of SORA -1

■ **As dramatic as the introduction of Chat-GPT:**

- The fear of making many image generation companies out-of-business
- Killer App for Tik-Tok?
- Many more dooms day views of visual generation tools and apps, and jobs..

■ **Structure of LFM:**

- Tokenizer/ Encoder: from latent representation -> (space-time) tokens
- Cross-modality Alignment (training) & Transfer (inference):
Architecture: U-net vs. Transformer, or combination (DiT: Diffusion Transformer)
Model: Diffusion vs. Autoregressive
- De-tokenizer/ Decoder: latent tokens -> image/video

The Emergence of SORA -2

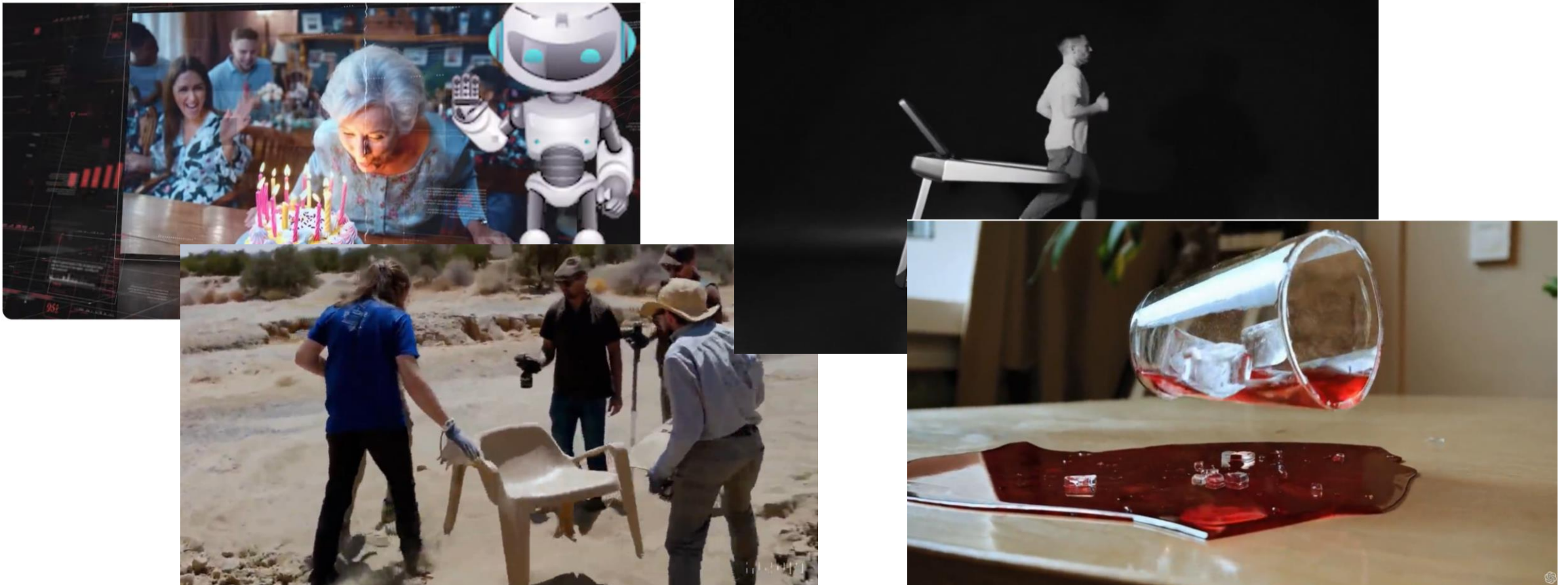
■ **Sora:**

- Believed to be adapting DiT architecture
- Trained on >1B images and >several millions of hours of videos mostly from Youtube and games; >10B parameters
- Make extensive use of data augmentation to generate high quality and more detailed captions for images and videos using Dalle-3, of higher quality and more consistent than human generated annotation
- Leverage (frozen) LLM to guide the generation of videos

The Emergence of SORA -3

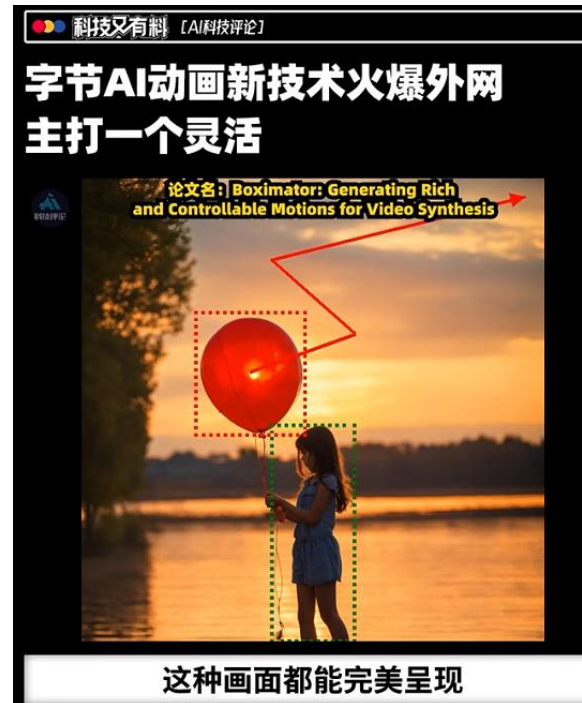
- **Key limitations:** Transformer is good at learning local token relations, but fails at distance relations. Hence individual objects and motions are generated with high quality, but fails in obeying the laws of physics in overall environment.

<Google search: videos sora openai



Key Research Directions

- Image/video generation with constraints: artistic rules, symmetry, enforce numerical and positional constraints.
- Extend DiT to learn distant relations between tokens: incorporate knowledge and physical laws into generation model
- Video editing makes simple: the power of Tiktok is in making video editing simple and natural



Requirements for Brave-New-Idea (BNI) papers

■ **AIM of BNI Paper:**

- 1) To propose a work that contain original ideas and research vision.
- 2) The paper should offer: (i) novel, exploratory solutions with sufficient evidence of proof-of-concept; (ii) visions describing a new or open problem in multimedia research; and/or (iii) a novel perspective on existing multimedia research.

■ **Guidelines:**

- Must be in multimedia and is expected to have a high component of novelty
- Should address an understudied, open problem in multimedia, while the ideas should be supported with sufficient scientific argumentation, experimentation and/or proof.
- The paper should contain ideas not previously submitted nor published.
- Should be within **5 pages**, excluding references, in ACM 2-column format.

■ **Key Deadlines for BMI Papers:**

- **Have received all abstracts.** Will feedback to you by the end of this week.
- Final paper Due: 5 Apr (Fri) @ 1700
- Presentation to Class (5 mins each): 9 Apr (1100-1200) & 16 Apr (1000-1200)

Short Idea/ Opinion 2

■ Topic:

Trust and Robustness in LFM (Large Foundation Models)

■ Outline of Paper:

- Robust and trust are the key problems to LFM. With the development of more powerful LFM with strong generative capability, this problem is becoming more severe.
- What are the key issues here. Are these fundamentally unsolved problems? What can be done to address the problems?
- What guidelines should be in place to mitigate such problems.
- The article should be within **3 pages**, in ACM 2-column format (excluding references).

■ Grading Guidelines:

- I am looking for new angles into the issues, as well as innovative ideas, insights and solutions.
- I will award a **B** if the paper covers most points above, and **A** for innovative ideas and insightful solution.

■ Deadlines:

- Article 2: 11 March @1700 (Submit-Article-2)

Papers for Lecture 8 (Large Multimodal Foundation Model)

- P8-1: Large Visual Foundation Models: Presenter: Qin Hangyu; Reader: He Yingzhi**
(Must-Read) H Liu, C Li, Q Wu, et al. Visual Instruction Tuning. NeurIPS 2023.
(SOTA) R Dong, et al. DreamLLM: Synergistic Multimodal Comprehension and Creation. ICLR 2024.
(Must-Read) D Zhu, et al. MiniGPT-4: Enhancing Vision-Language Understanding with Advanced LLMs. ICLR 2024.
(Background) H Touvron, et al. LLaMA: Open and Efficient Foundation Language Models. arXiv 2023.
(Background) T Brown, et al. Language Models are Few-Shot Learners. arXiv 2022.
- P8-2: Pixel Grounding Large Multimodal Models: Presenter: Stefan Putra Lionar; Reader: Cao Xiao**
(SOTA) Y. Yuan, et al. Osprey: Pixel Understanding with Visual Instruction Tuning. arXiv 2023.
(Must-Read) H. Rasheed, et al. Glamm: Pixel grounding large multimodal model. arXiv 2023.
(To-Read) Z. Ren, et al. PixelLM: Pixel Reasoning with Large Multimodal Model. arXiv 2023.