Lecture 6:

Image Video QA, and Reasoning

Papers for Lecture 6 (Image/ Video QA, and Reasoning)

P6-1: Image QA and Reasoning: Presenter: Wu Yihang; Reader: Cheng Yi

(Classic SOTA-1) P Anderson, X He, C Buehler, D Teney, M Johnson, S Gould, & L Zhang. Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering. CVPR 2018.

(Classic SOTA-2) J Lu, J Yang, D Batra, et al. Hierarchical Question-Image Co-Attention for Visual Question Answering. NeurIPS 2016.

(Popular Dataset, To-Read) D A Hudson & C D Manning. GQA: A New Dataset for Real-World Visual Reasoning and Compositional Question Answering. CVPR 2019.

P6-2: Video QA and Reasoning: Presenter: He Yingzhi; Reader: Yannis Mohamed Christian Montreuil

(Must-Read) A J Piergiovanni, K Morton, W Kuo, et al. Video Question Answering with Iterative Video-Text Co-Tokenization. ECCV 2022.

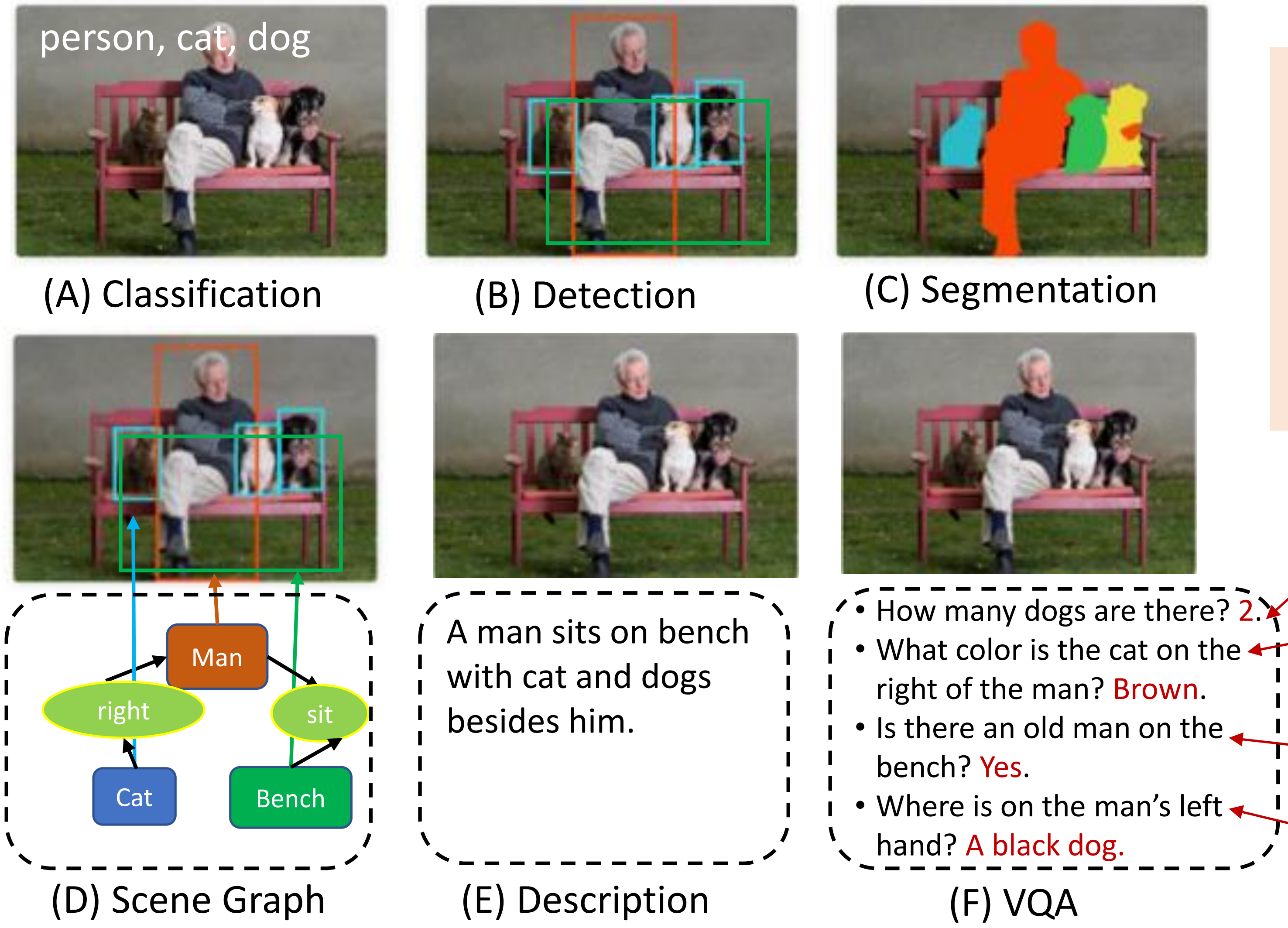
(Must-Read) A Yang, A Miech, J Sivic, I Laptev & C Schmid. Just Ask: Learning to Answer Questions from Millions of Narrated Videos. CVPR 2021.

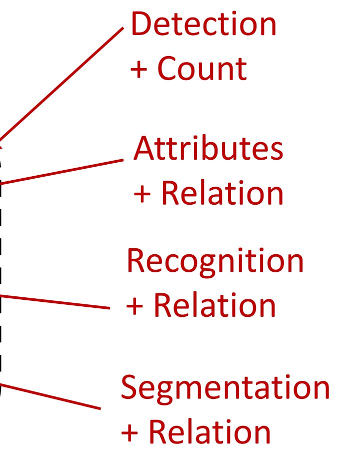
(Survey: To-Read): Y Zhong, W Ji, J Xiao, Y Li, W Deng & TS Chua. Video Question Answering: Datasets, Algorithms and Challenges. EMNLP 2022.

(First Dataset, Must-Read): J Xiao, X Shang, A Yao & TS Chua. NExT-QA: Next Phase of Question-Answering to Explaining Temporal Actions. CVPR 2021

From Visual Classification to Captioning and QA

- VQA, towards AI-compete visual understanding





- These are mostly factoid QA
- Like text-QA, the main challenges are to go towards temporal, reasoning-based and generative QA

VQA Definition

- VQA aims to leverage advanced visual and linguistic analytics, as well as external knowledge, to provide precise answers to user's natural language questions about the visual contents [1,2].

Problem Formulation:

- **Input**: Video, Question.
- **Output**: (Video) Answer
- **Task Settings**: Classification (Multiple-choice & Open-ended)

$$a^* = \arg\max_{\{a \in A\}} P(a \mid v, q;\ \theta)$$

What was behind the lady when she was bent down?
A0 A painting
A1 A couch
A2 A metal shelf
A3 A file cabinet
A4 A car

How many times does the cat touch the dog? 4
What action does the cat do 4 times? Touch dog

[1] R. Hong, G. Li, L. Nie, J. Tang, & T.-S. Chua, "Exploring large scale data for multimedia qa: an initial study," in CIVR. ACM, 2010, pp. 74–81.
[2] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. Lawrence Zitnick, and D. Parikh, "Vqa: Visual question answering," in ICCV, 2015, pp. 2425–2433.

Types of QA and VQA

- Text-based QA has been a hot research topic since 2000.
- Evolution of text-based QA:
  - Factoid QA
  - Knowledge-based QA, Definition QA, List QA
  - Reasoning (what-if), Temporal, Opinion QA etc.
  - Interactive QA & Dialogue
- Visual QA follows the same evolution path:
  - Factoid VQA: most current systems
  - Temporal and reasoning VQA – current trends
  - Multimodal Dialogue – emerging development, will accelerate following ChatGPT

Image and Video QA

ImageQA

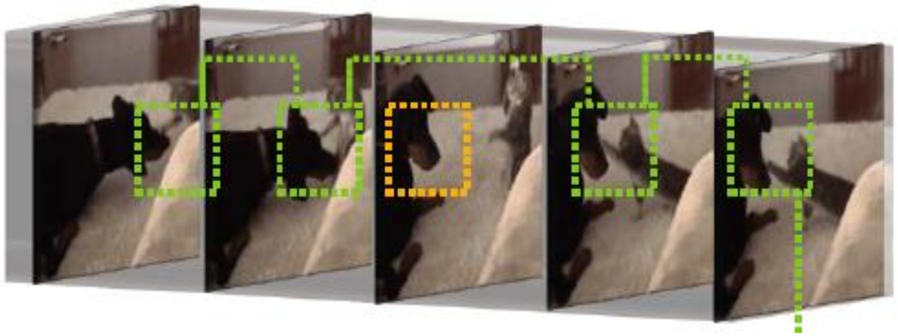
(Example from Antol *et al*. ICCV 2015)

What color is her mustache? Yellow
What is the mustache made of? Banana

Can I use this to buy something? Yes

VideoQA

(Example from Jang *et al*. CVPR 2017, and Xiao et al. 2021)

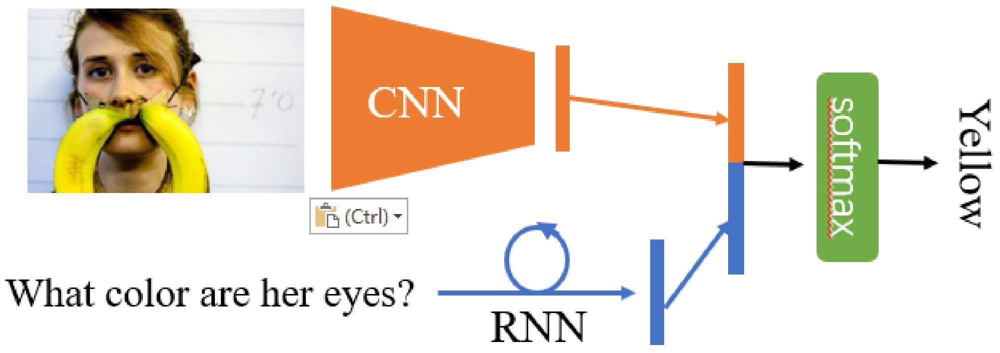


How many times did the cat touch the dog? 4
What action does the cat do 4 times? Touch dog

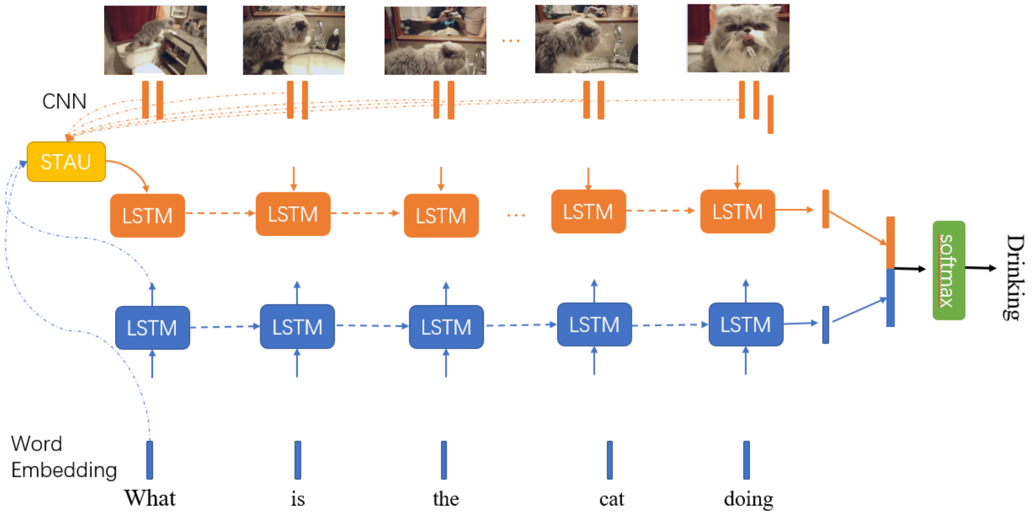
Why did the toddler in red cry at the end of the video? Fell backwards.

VQA Methods

ImageQA:



VideoQA:



Visual-Text Interaction

- Hierarchical attention (word/ phrase/ sentence)
- Co-attention (img2qns, qns2img)
- Stack attention (multi-turn).
- Bottom-up & top-down attention (region proposals & partially-completed sequence output)

Also incorporate additional info

Video Embedding

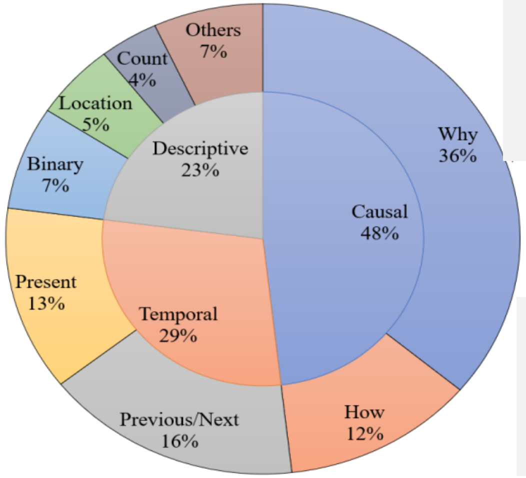
- Spatial-temporal attention.
- Appearance & motion feature.
- Multi-granularity (frame/clip/video).

A Study of SOTA Methods on Temporal & Causal Questions

Analysis of performance:

- **NExT-QA dataset:** 48% causal(C), 29% temporal (T) and 23% descriptive (D) questions
- **Multiple-choice questions:** Performances are good (≥ 43% in accuracy); gap between C&T vs. D questions is ≈10%
- **Open questions: P**erformances are bad for SOTA methods, especially for C&T questions, marginally better than blindQA; gap between C & T vs. D questions is > 30%

- Current SOTA methods are weak on C&T questions , and do not have good understanding of language & visual content
- **Recent approaches formulate VQA as a ranking problem**



(a) Distribution of question types

Methods	Acc_C	Acc_T	Acc_D	Acc
EVQA [2]	43.27	46.93	45.62	44.92
STVQA [17]	45.51	47.57	54.59	47.64
CoMem [11]	45.85	**50.02**	54.38	48.54
HCRN [24]	47.07	49.27	54.02	48.89
HME [9]	46.76	48.89	57.37	49.16
HGA [19]	**48.13**	49.08	**57.79**	**50.01**

Table 5: Results of multi-choice QA on test set. All are based on fine-tuned BERT representation.

Methods	$WUPS_C$	$WUPS_T$	$WUPS_D$	$WUPS$
Popular	12.19	10.79	31.94	16.12
BlindQA	14.87	18.35	45.78	22.66
STVQA [17]	15.24	18.03	47.11	23.04
HCRN [24]	16.05	17.68	49.78	23.92
HME [9]	15.78	18.40	50.03	24.06
UATT [54]	16.73	**18.68**	48.42	24.25
HGA [19]	**17.98**	17.95	**50.84**	**25.18**

Table 7: Results of open-ended QA on test set. We provide two reference answers for half of the test questions, and report the highest WUPS score between them.

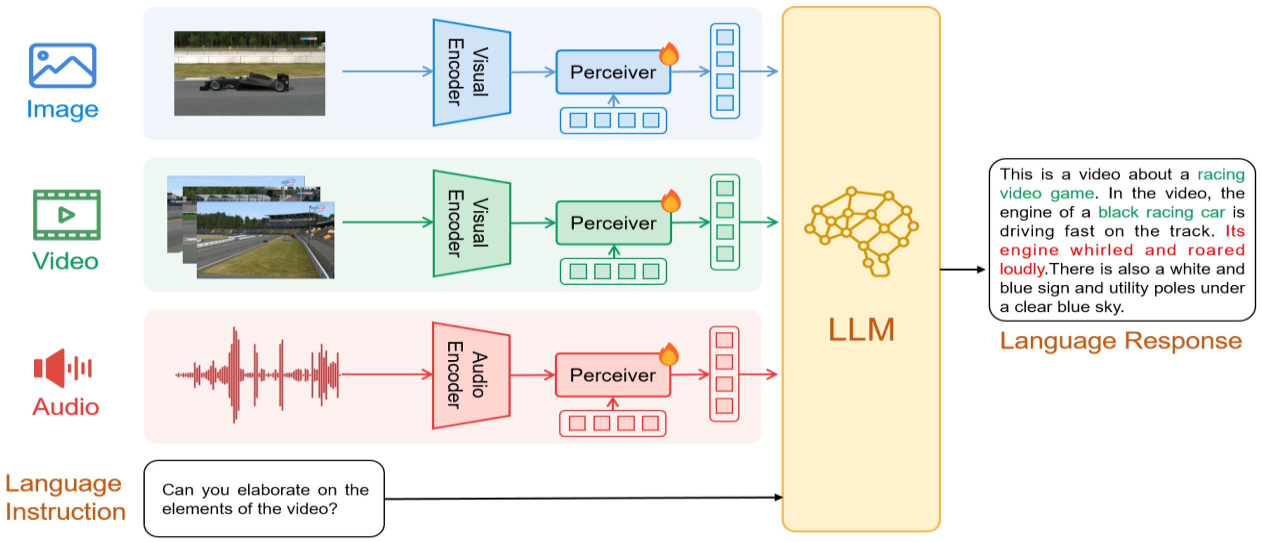
Xiao et al. NExT-QA: Next Phase of Question-Answering to Explaining Temporal Actions. CVPR 2021.

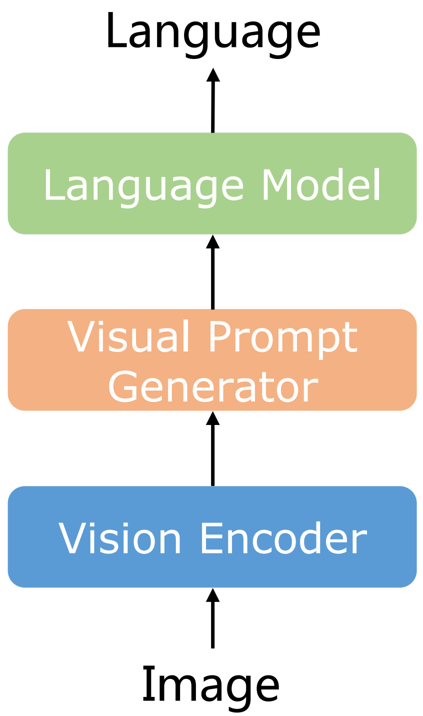
Current Trends in VQA research

The impact of LLM:

- Towards enhancing LLM with multimodal capabilities: Multimodal Foundation Model
  - Use LLM as the brain for Multimodal QA and Conversation
  - General idea: adapt frozen instruction-tuned LLMs to understand multimodal inputs and generate multimodal outputs
  - Move towards aligning text or image with multimodalities

  For example: ImageBind, ChatBridge, PandaGPT, NExT_GPT, etc.

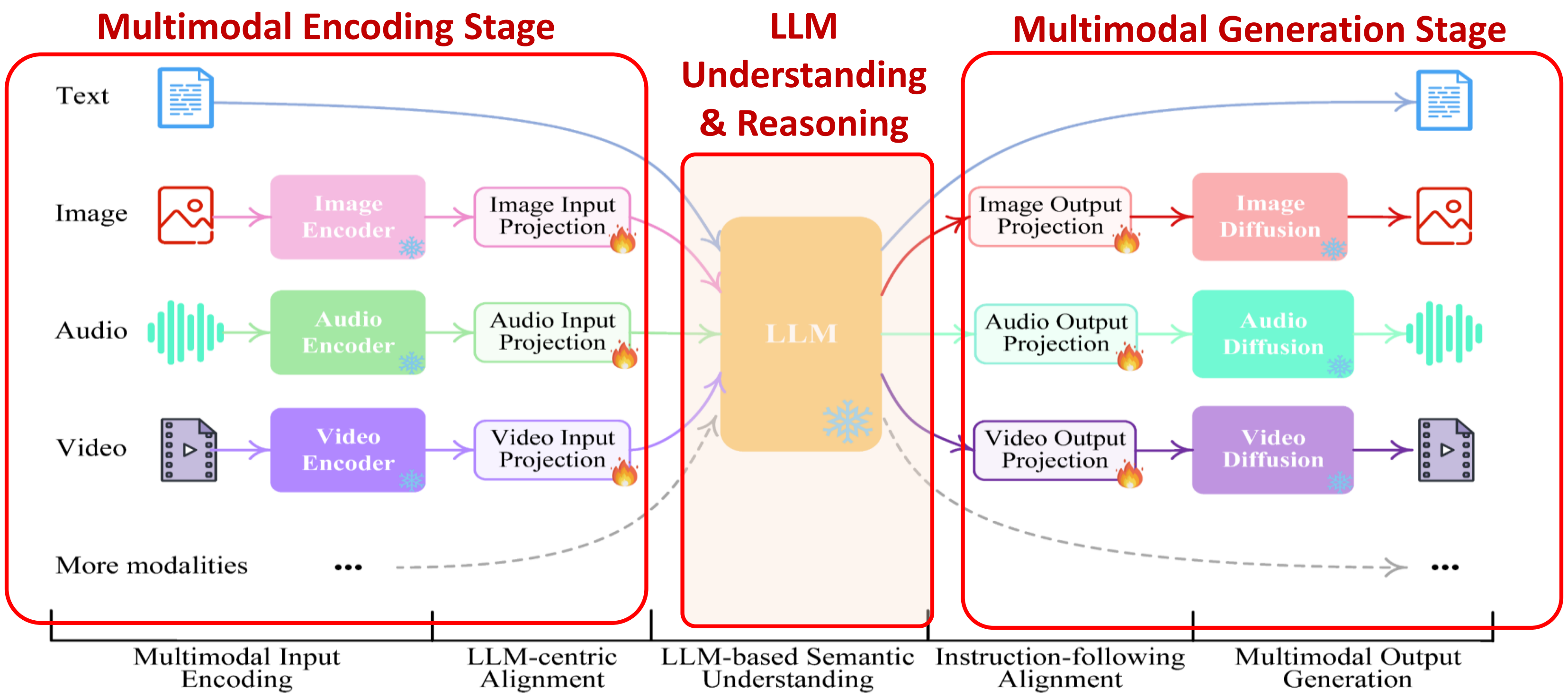




Current Trends in VQA research

The impact of LLM:

- LLM-Enhanced Multimodal Framework for QA and Conversation
  - Support multimodal input and multimodal output
  - LLM performs reasoning and determines the best media as answers (the generation)
  - **Many remaining challenges**



Project: https://next-gpt.github.io; Paper: https://arxiv.org/pdf/2309.05519

Short Idea/ Opinion 1

- **Topic:**

  Can LFM (Large Foundation Model) use public data for training and content generation: what are the issues and guidelines?

- **Requirements for the Paper:**

  Deadline: 16 Feb @1700

  Thanks all for submitting on time

- **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 1: 16 Feb @1700 (Submit-Article1)

Requirements for Brave-New-Idea (BNI) papers

- **Key Deadlines for BMI Papers:**
  - Submission of title and Abstract: 24 Feb (Sat), @Submit-BNI-Abstract
  - Final paper Due: 5 Apr (Fri) @ 1700
  - Presentation to Class (5 mins each): 9 Apr (1100-1200) & 16 Apr (1000-1200)

- **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.
- **Grading Criteria:**
  - Novelty; Conceptual leap; Depth of Impact; Breadth of impact

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.