2023-06-12
Better Figures

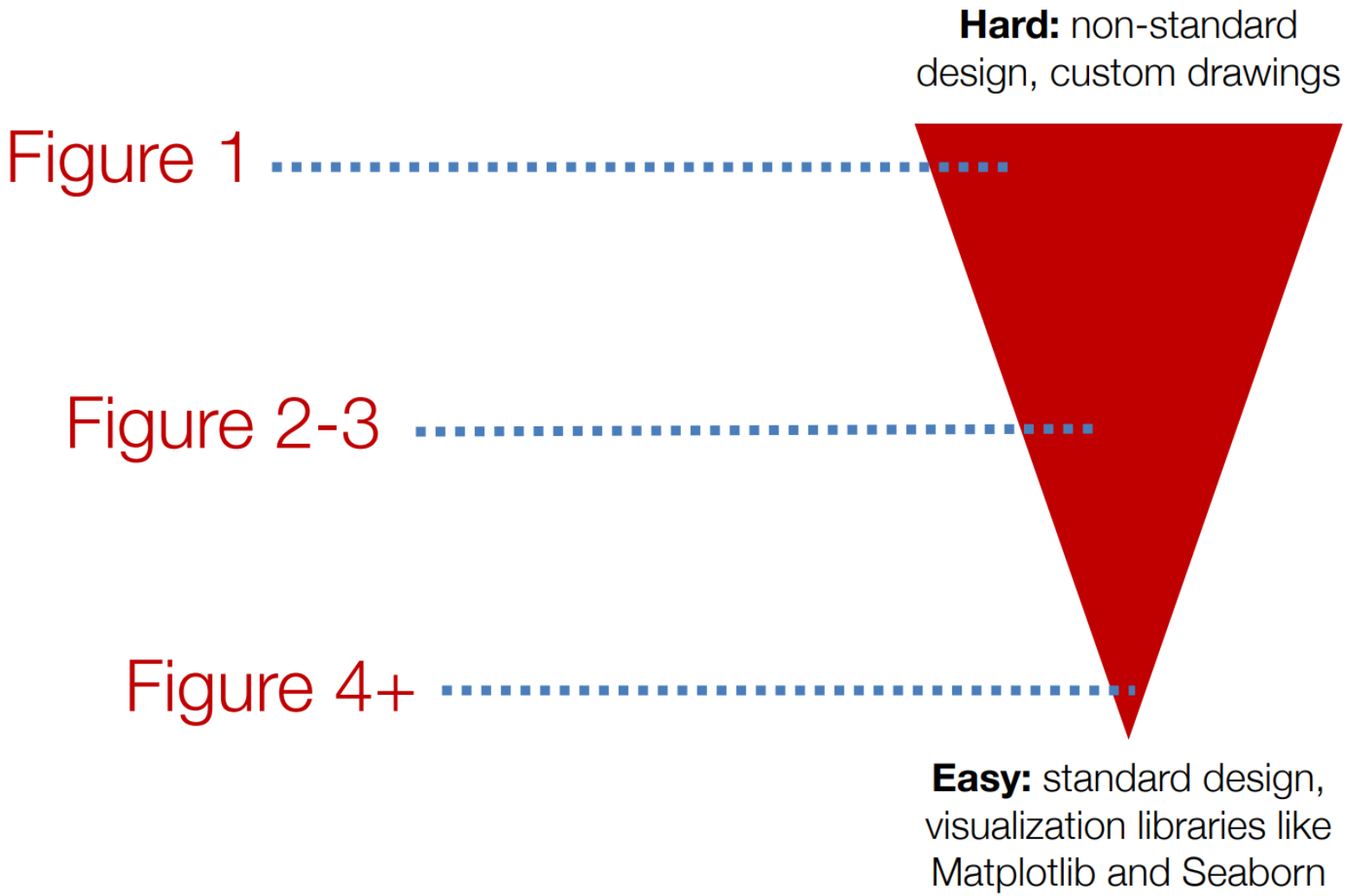
Core CS Conference Papers:

Visual Structure

NeurIPS, ICML, ICLR, AAAI, etc.

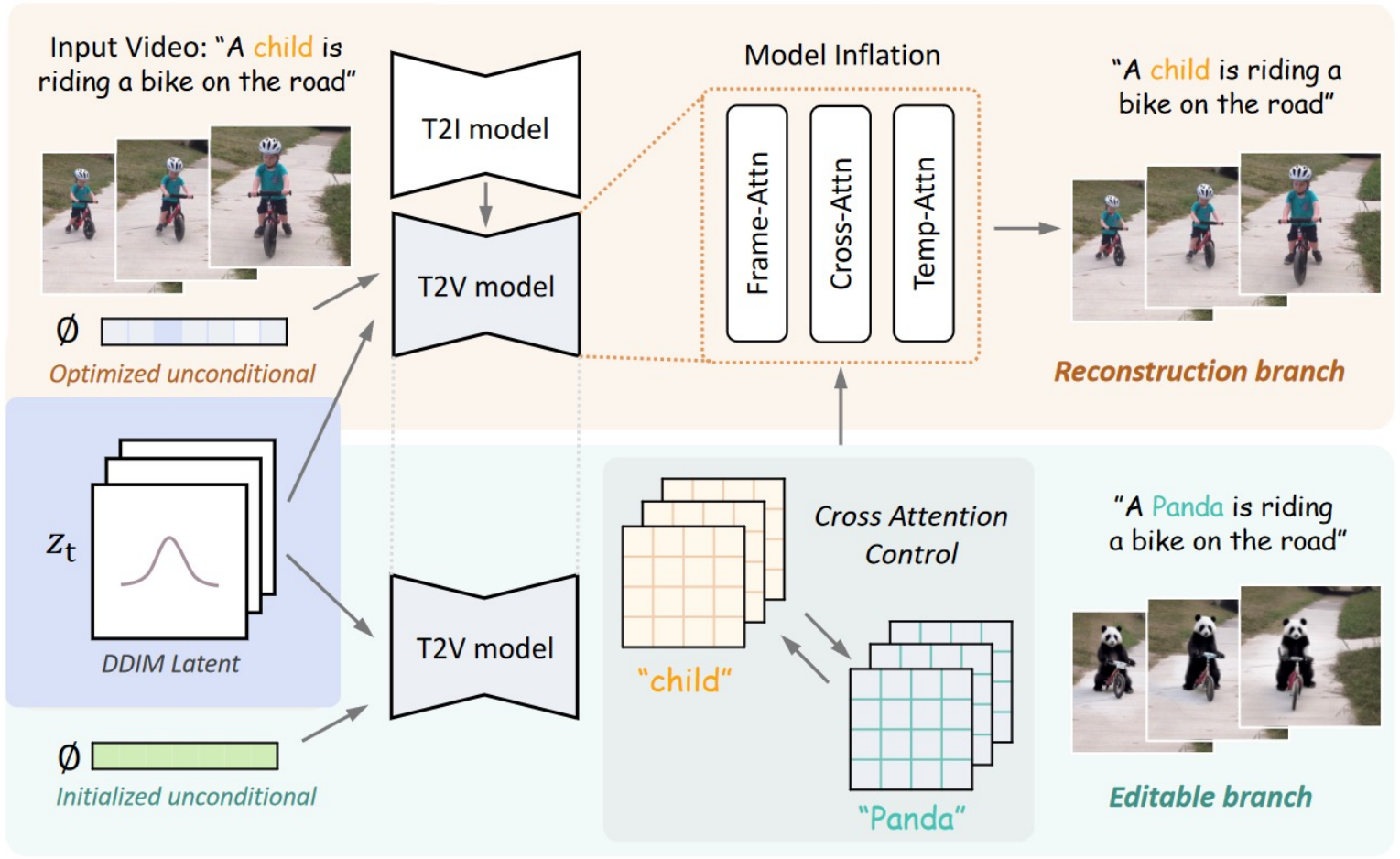
- Figure 1: Key methodological contribution
  - Focus on most important information
  - **Impress your audience!**
    - Is your method/system the fastest, the largest, the most accurate?
    - What is the hard problem that your method solves?
    - What makes your method different from related work?
- Figure 2-3: Overview and algorithmic details
  - Inputs + Data transformation + Outputs
  - Show details about data transformations:
    - Graph convolutions, neural architectures, etc.
- Figure 4+: Results

Core CS Conference Papers:
Visual Structure



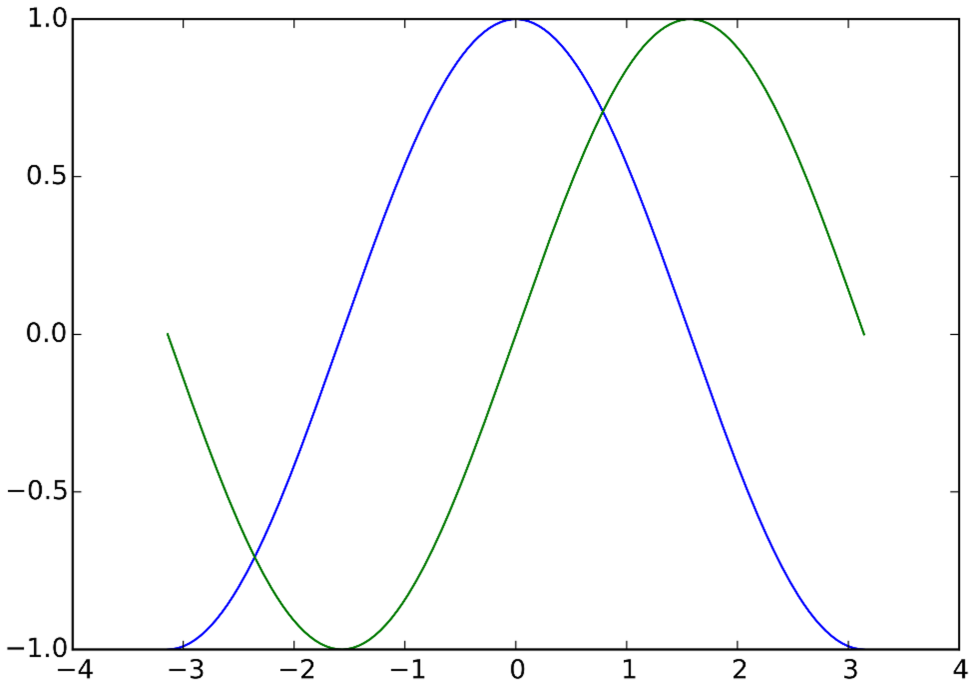
1 Design figures for the audience (not for you)

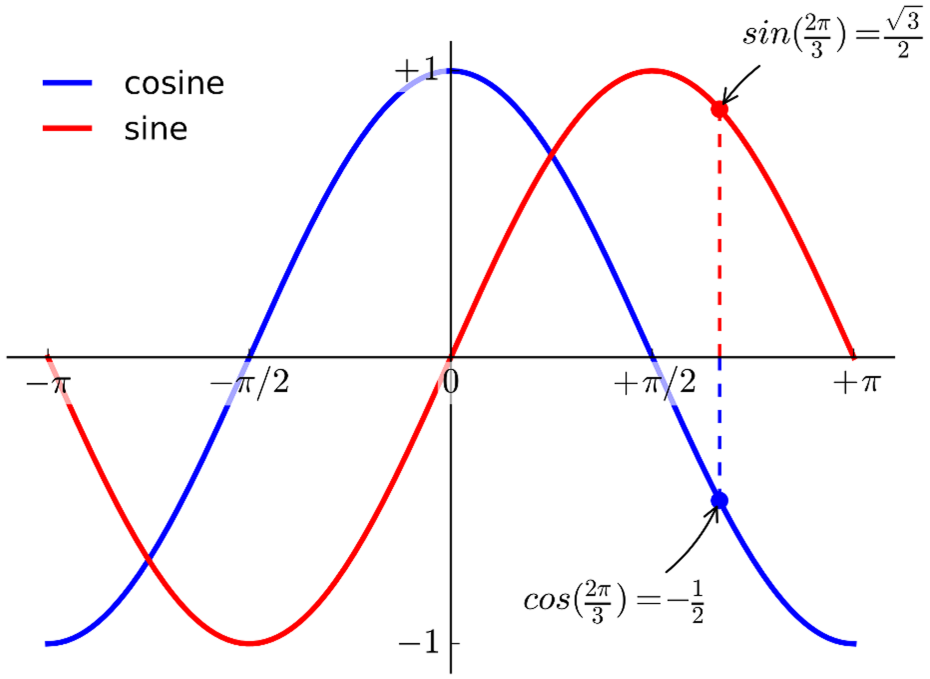
- Make-up of the audience
  - Will a figure appear in a specialized journal?
  - Is a figure aimed at a broad readership?
- Background knowledge of the audience
  - Audience may not know what you know
  - Figures should provide all the information necessary for the audience to fully comprehend them


Input Video: "A child is riding a bike on the road"
T2I model
T2V model
Model Inflation
Frame-Attn
Cross-Attn
Temp-Attn
"A child is riding a bike on the road"
Ø
Optimized unconditional
Reconstruction branch
z_t
DDIM Latent
T2V model
Cross Attention Control
"child"
"Panda"
"A Panda is riding a bike on the road"
Ø
Initialized unconditional
Editable branch

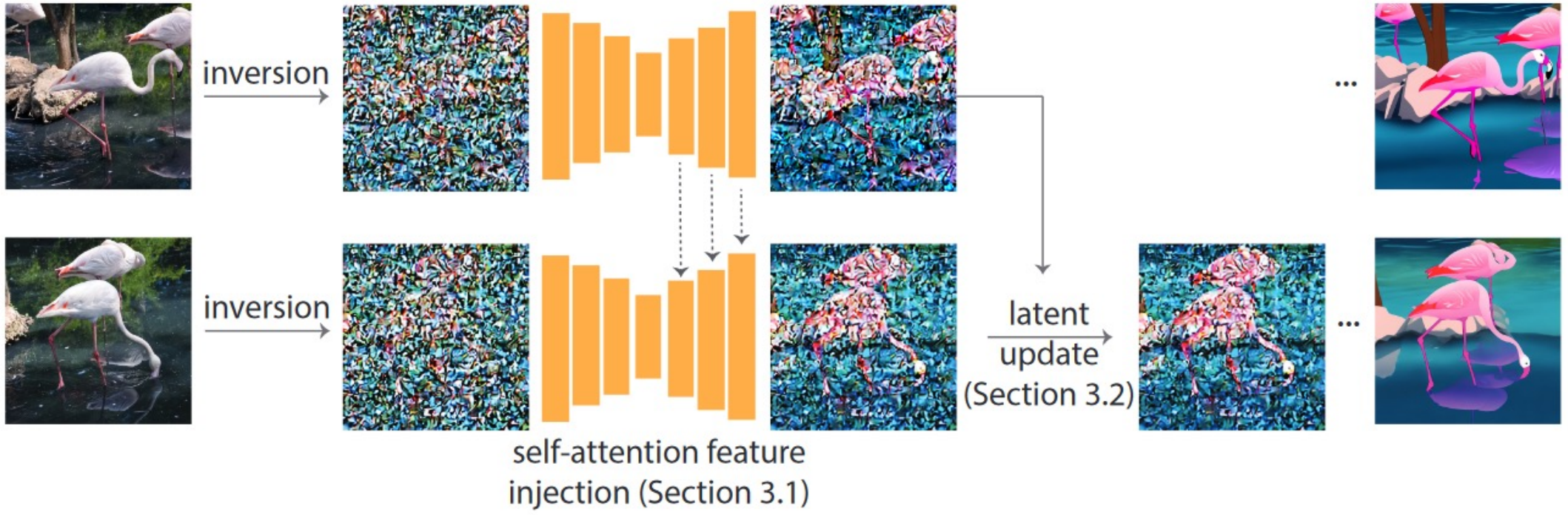
Video-p2p

2 Identify Your Message

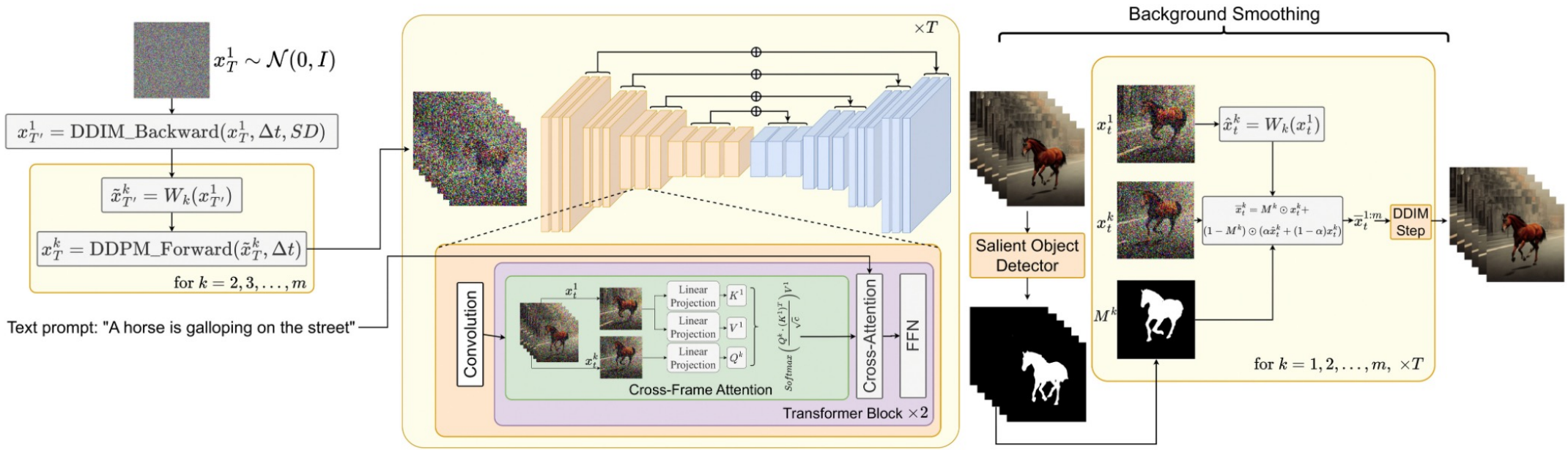




Message Trumps Beauty

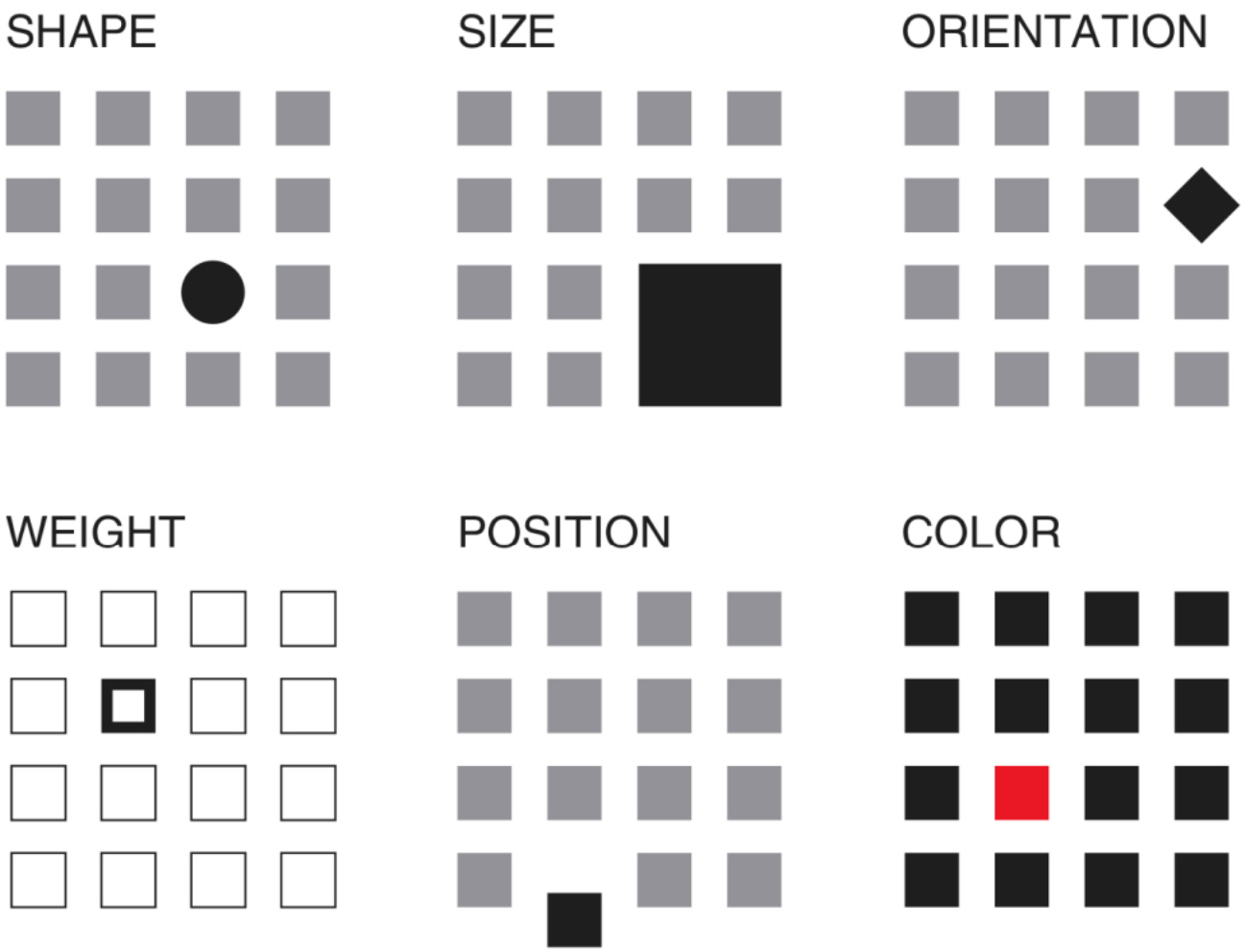


pix2video

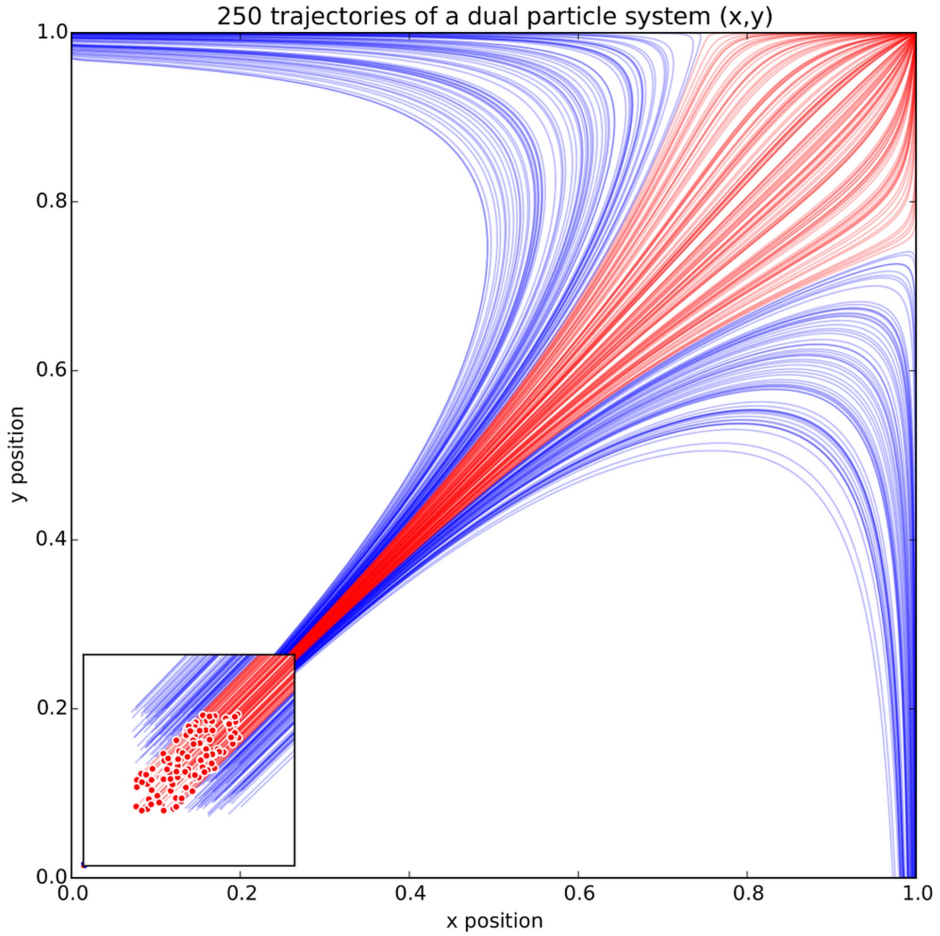


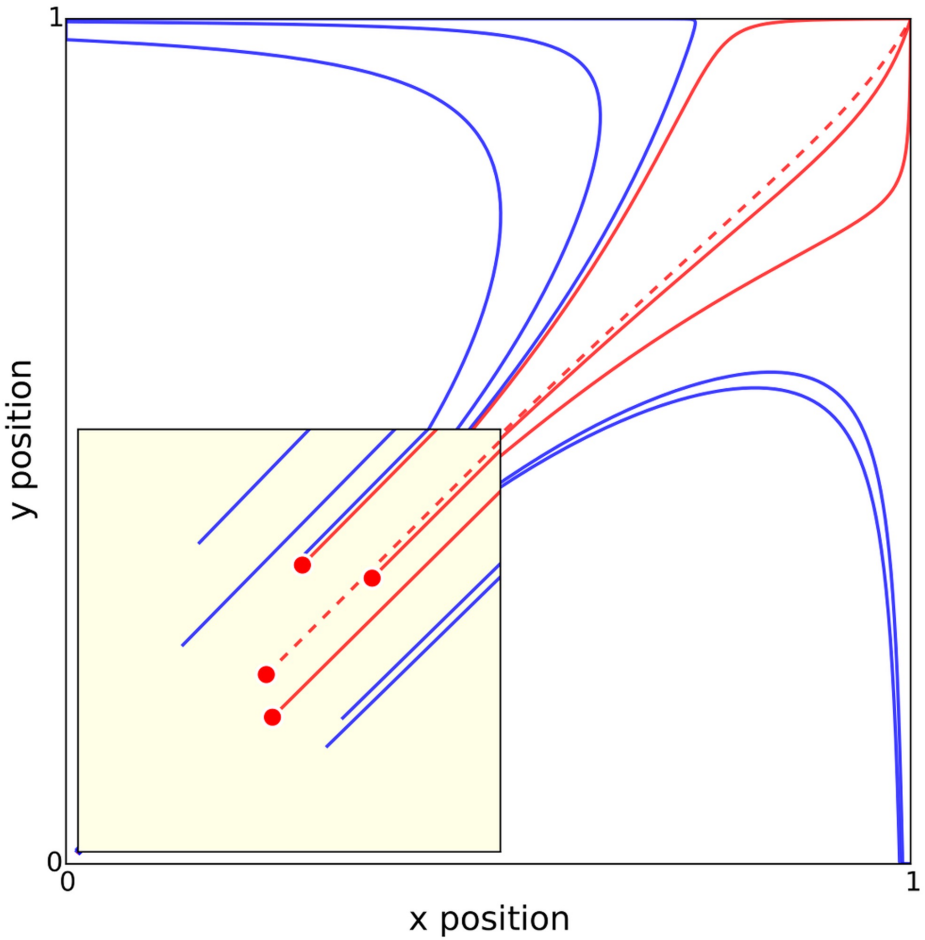
Text2video-zero

3 Use visual contrast



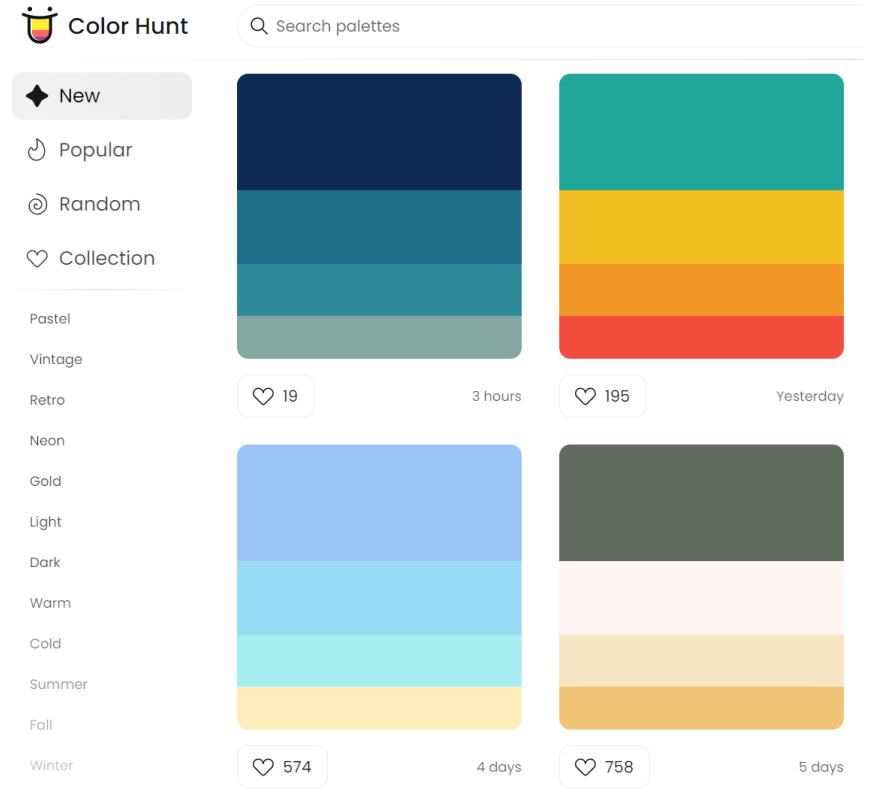
Details

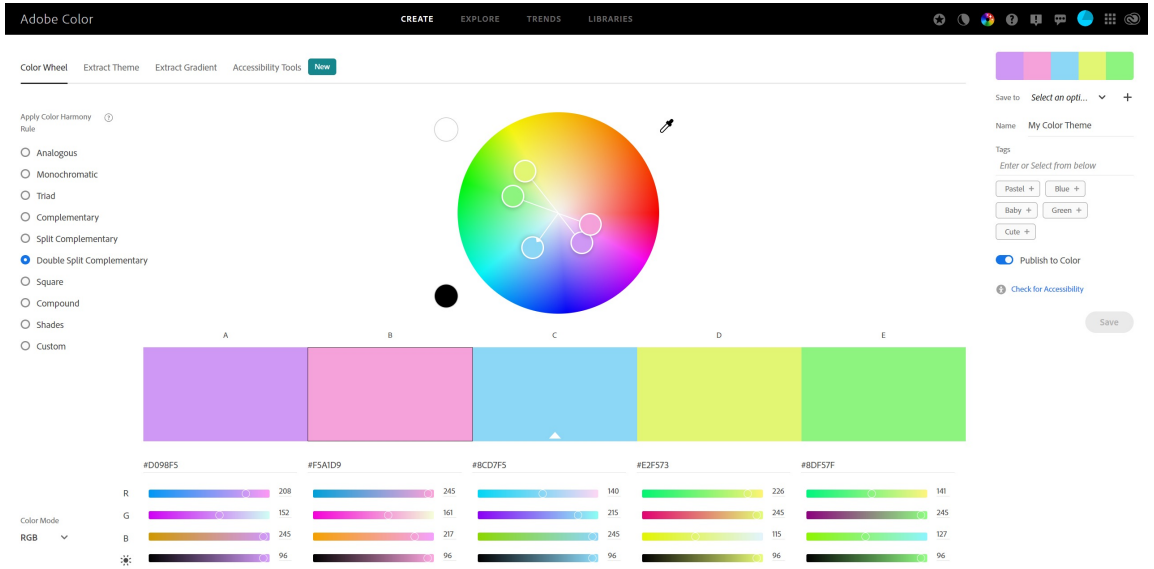

250 trajectories of a dual particle system (x,y)
y position
x position
0.0
0.2
0.4
0.6
0.8
1.0


y position
x position
0
1

Posters / illustrators

- ColorHunt: https://colorhunt.co/
- Pinterest: https://www.pinterest.com/
- Adobe color wheel: https://color.adobe.com/create/color-wheel

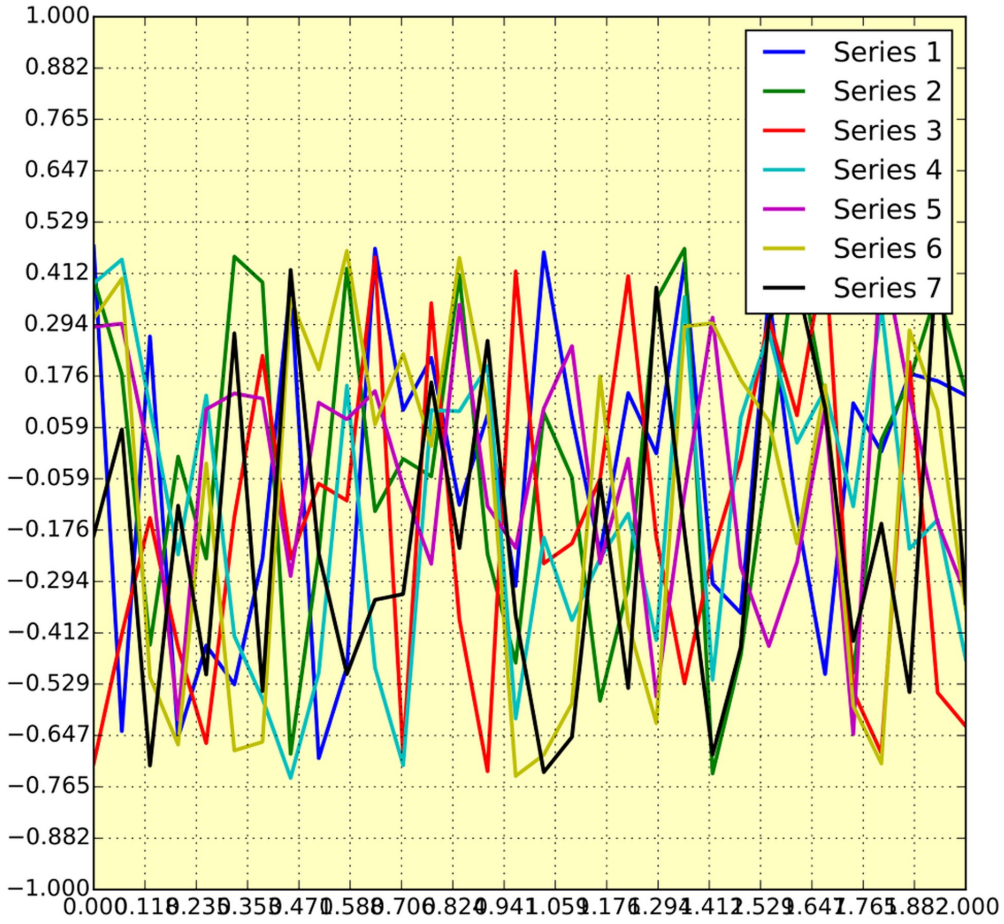


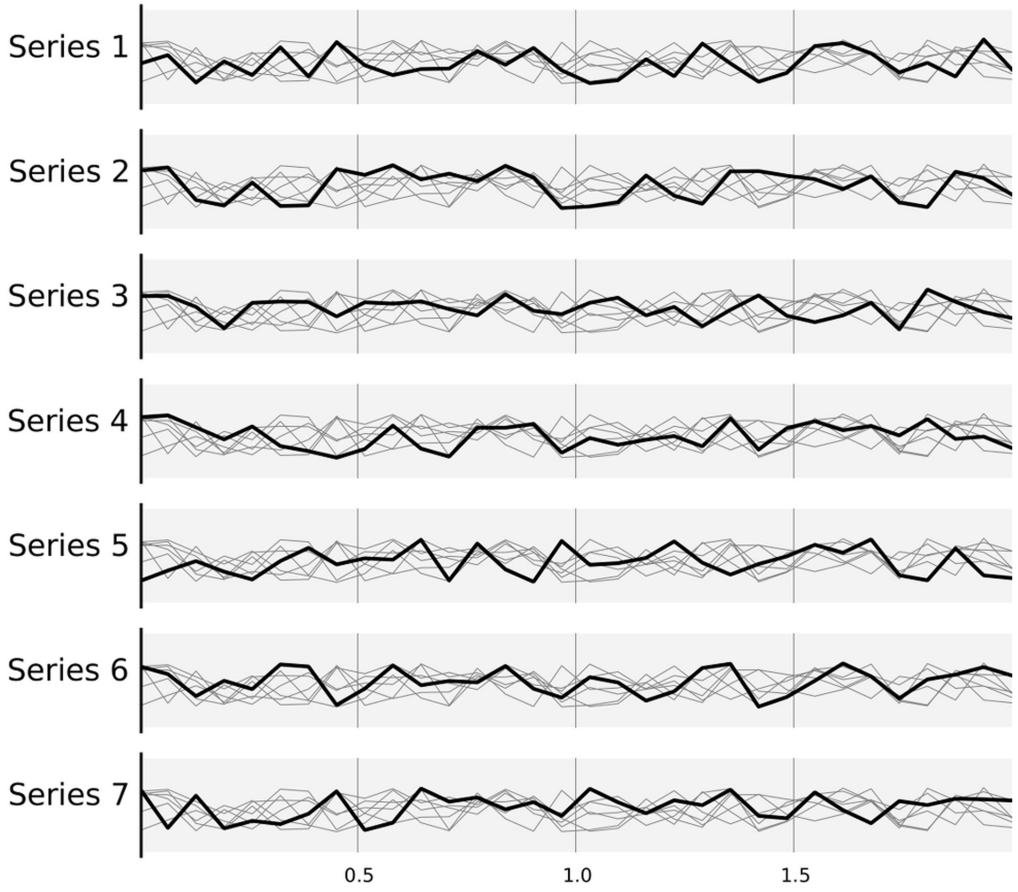


Icons / Fonts

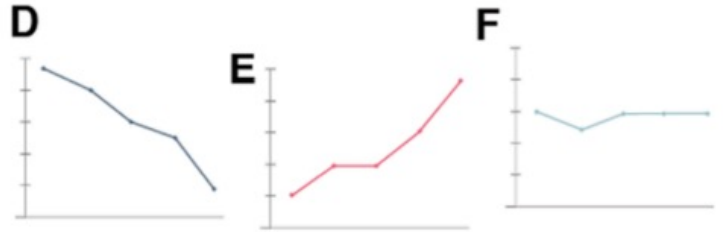
- Iconfinder: https://www.iconfinder.com/
- 阿里iconfont: https://www.iconfont.cn/
- Fontspace: https://www.fontspace.com/

4 Avoid “Chartjunk”, be consistent

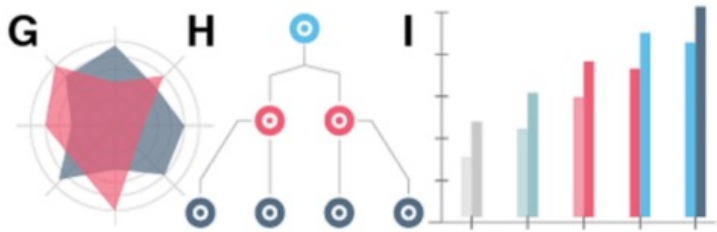




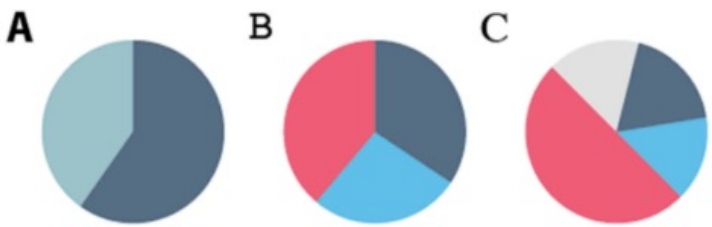
Lack of alignment



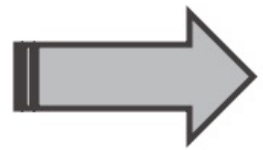
Insufficient padding



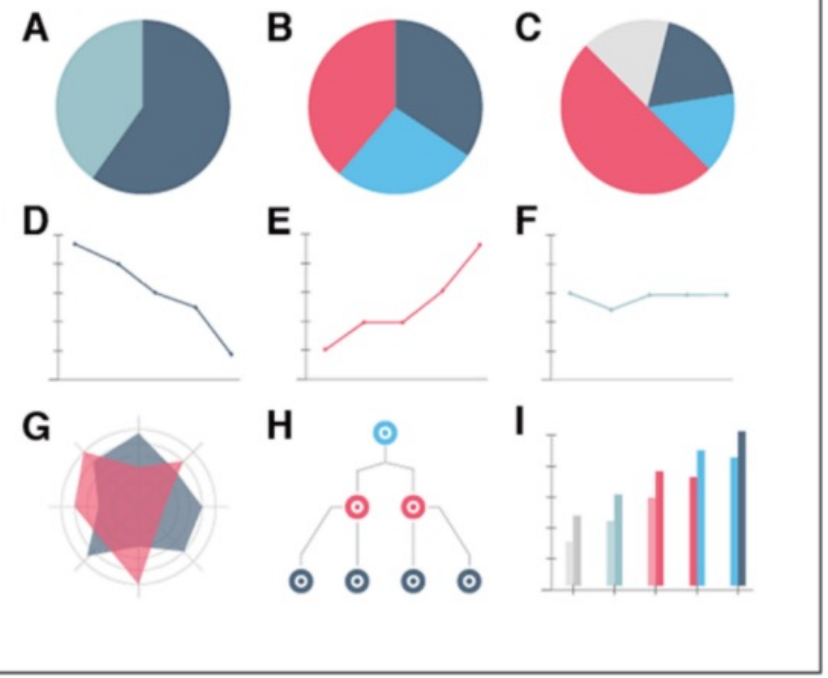
Inconsistent font



Source: Jean Fan, Harvard

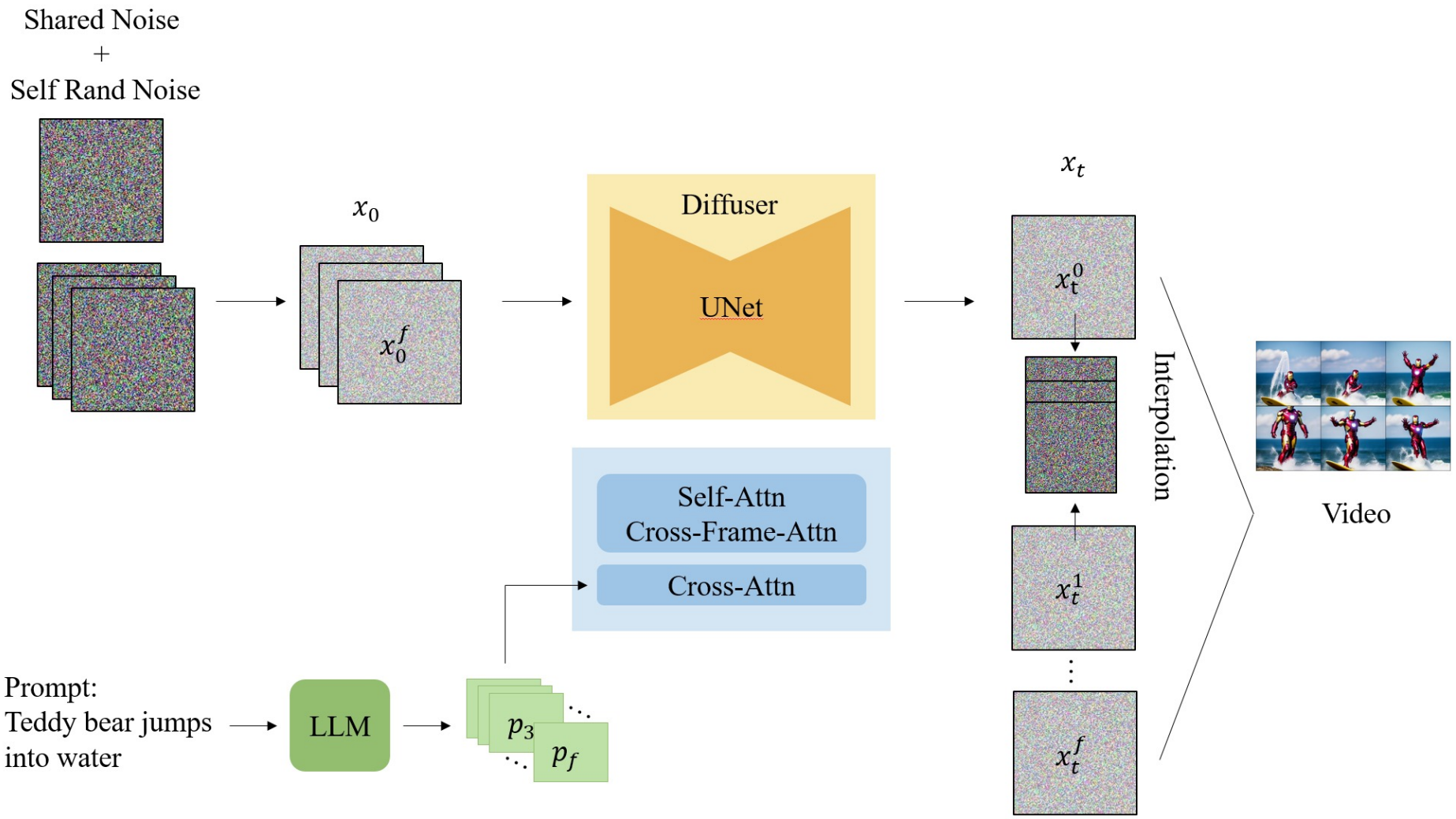


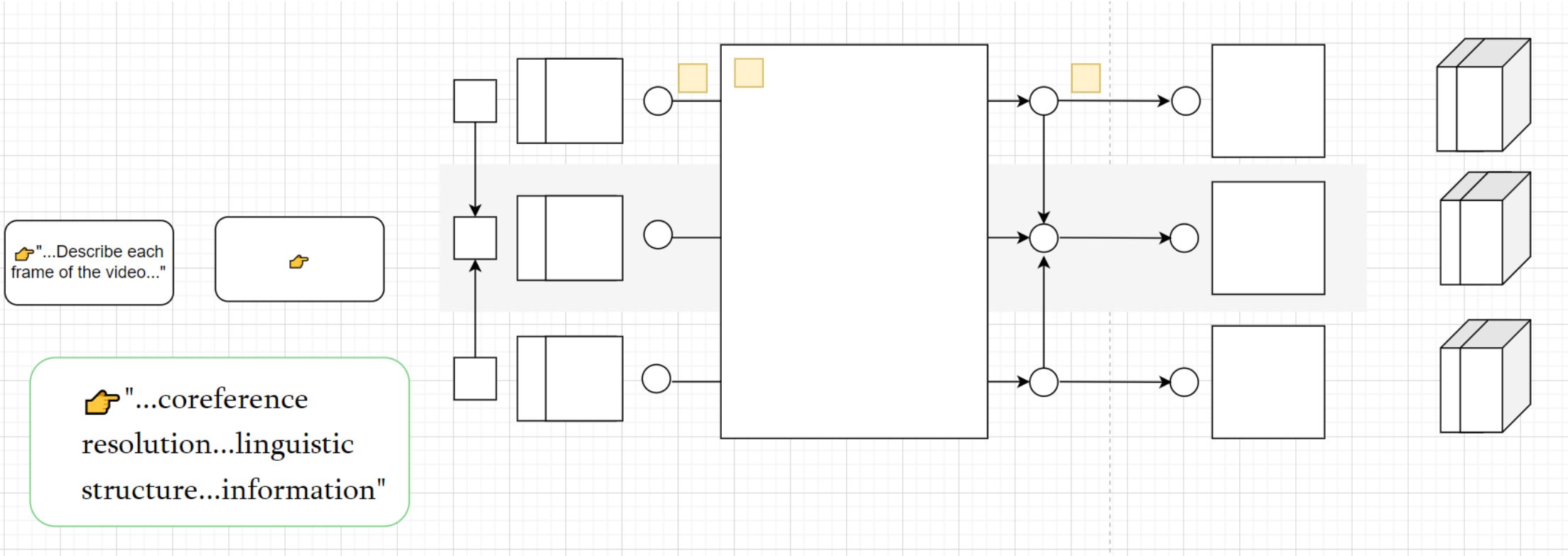
Fixed

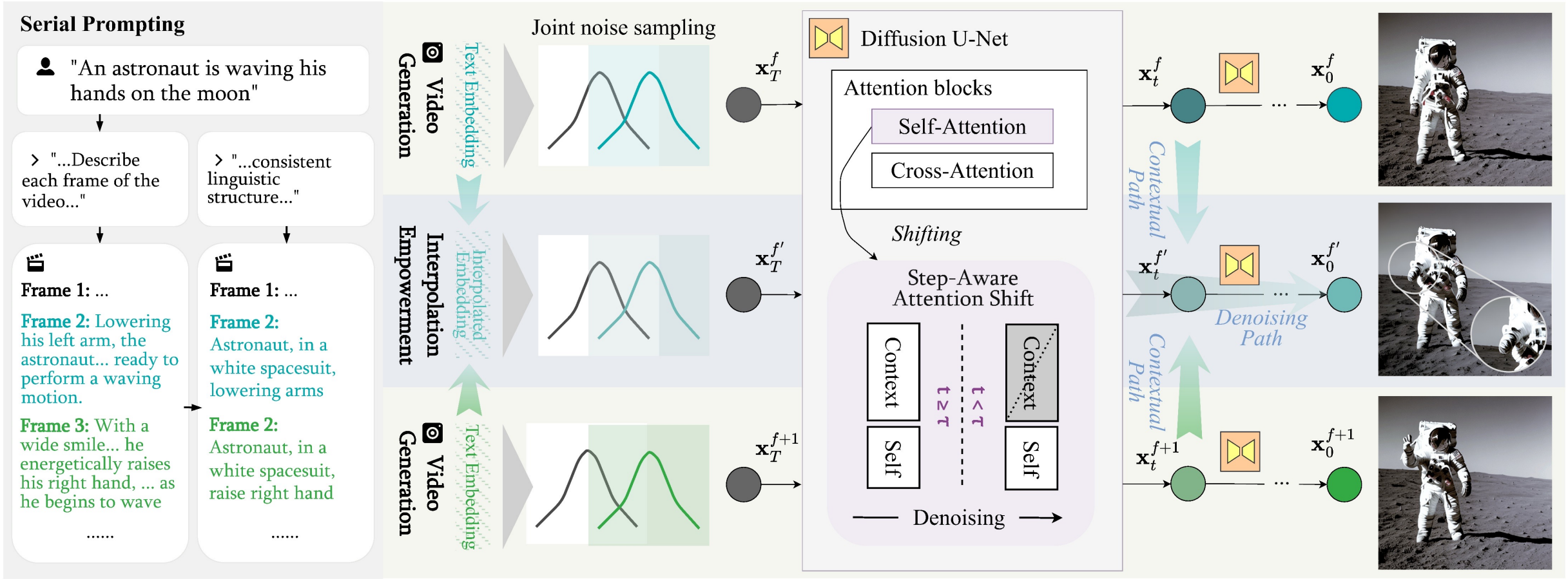


5 Captions Are Not Optional

- can be thought of as the explanation you would give during an oral presentation, or in front of a poster, but with the difference that you must think in advance about the questions people would ask.
- make sure it is visually distinct but do not hesitate to point it out again in the caption.


Shared Noise
+
Self Rand Noise
x_0
x_0^f
Diffuser
UNet
x_t
x_t^0
x_t^1
x_t^f
Interpolation
Video
Self-Attn
Cross-Frame-Attn
Cross-Attn
Prompt:
Teddy bear jumps
into water
LLM
p_3
p_f


👉"...Describe each frame of the video..."
👉
👉"...coreference resolution...linguistic structure...information"


Serial Prompting
"An astronaut is waving his hands on the moon"
> "...Describe each frame of the video..."
> "...consistent linguistic structure..."
Frame 1: ...
Frame 2: Lowering his left arm, the astronaut... ready to perform a waving motion.
Frame 3: With a wide smile... he energetically raises his right hand, ... as he begins to wave
.....
Frame 1: ...
Frame 2: Astronaut, in a white spacesuit, lowering arms
Frame 2: Astronaut, in a white spacesuit, raise right hand
.....
Video Generation
Text Embedding
Interpolation Empowerment
Interpolated Embedding
Video Generation
Text Embedding
Joint noise sampling
$\mathbf{x}_T^f$
$\mathbf{x}_T^{f'}$
$\mathbf{x}_T^{f+1}$
Diffusion U-Net
Attention blocks
Self-Attention
Cross-Attention
Shifting
Step-Aware Attention Shift
Context
Self
$t \geq \tau$
$t < \tau$
Context
Self
Denoising
$\mathbf{x}_t^f$
$\mathbf{x}_0^f$
$\mathbf{x}_t^{f'}$
$\mathbf{x}_0^{f'}$
$\mathbf{x}_t^{f+1}$
$\mathbf{x}_0^{f+1}$
Contextual Path
Denoising Path
Contextual Path


SEPT. 28–30
ENVIRONMENTS

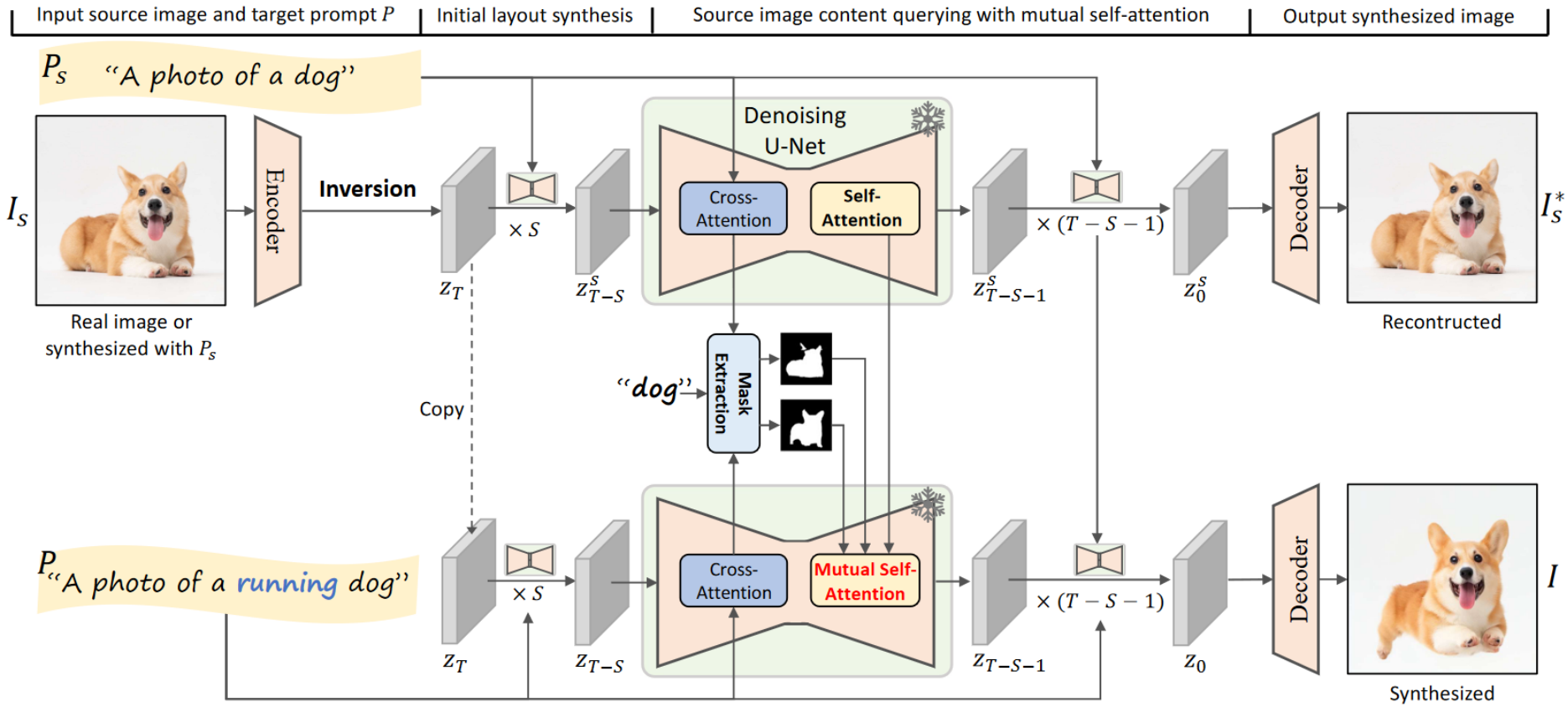


Figure 3: Pipeline of the proposed MasaCtrl. Our method tries to perform complex non-rigid image editing and synthesize content-consistent images. The source image is either real or synthesized with source text prompt P_s. During the denoising process for image synthesis, we convert the self-attention into mutual self-attention to query image contents from source image I_s, so that we can synthesize content-consist images under the modified target prompt P.

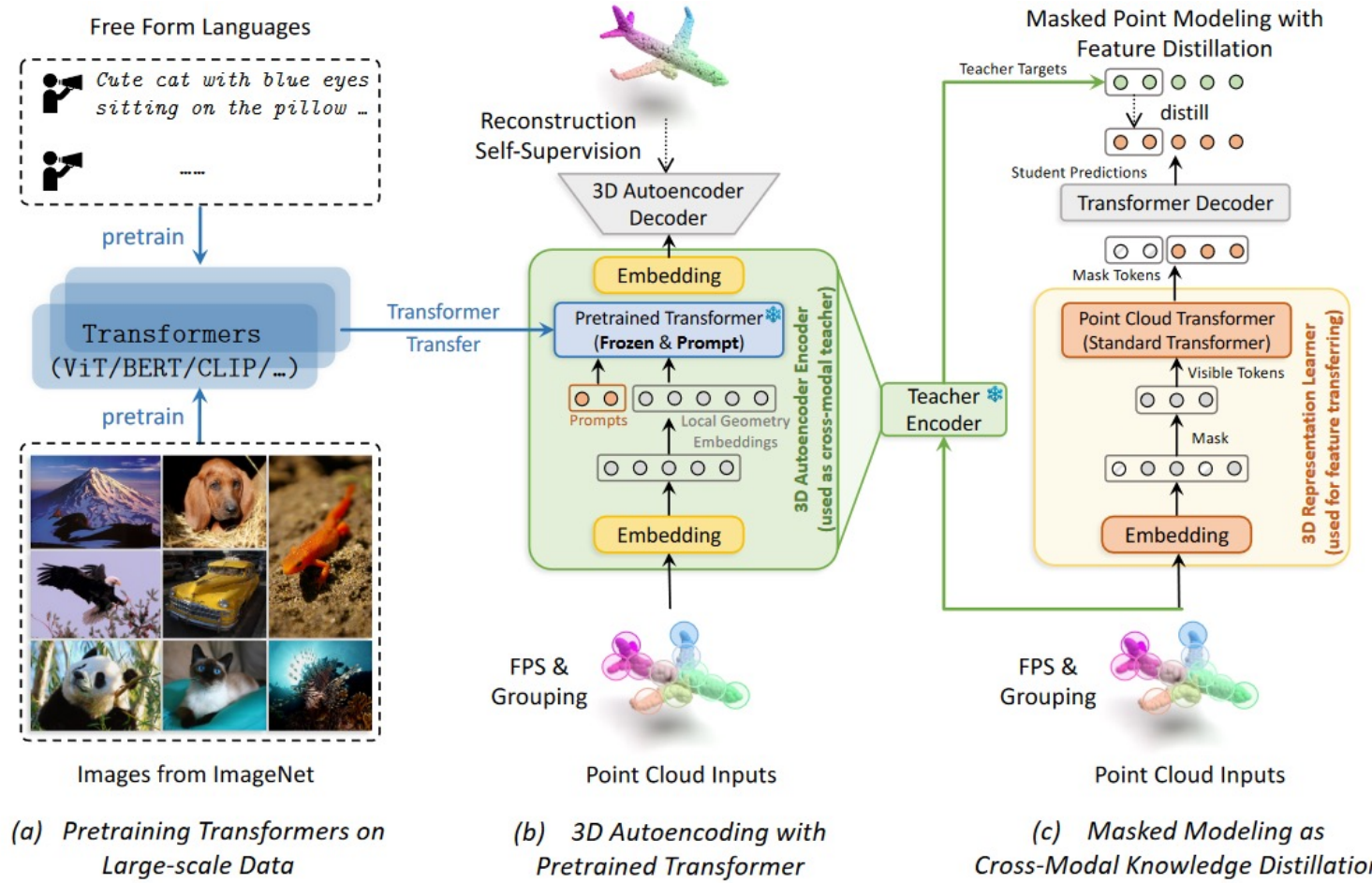


Figure 1: Overview of of our ACT framework (Sec. 3-4). (a) ACT utilizes the Transformers pretrained on large-scale data, *e.g.*, ViT (Dosovitskiy et al., 2021) pretrained with 2D images or BERT (Devlin et al., 2019) pretrained with languages. (b) Stage I of ACT (Sec. 4.1), the pre-trained Transformers are tuned by self-supervised 3D autoencoding with prompts (Jia et al., 2022). (c) Stage II of ACT (Sec. 4.2), the 3D autoencoder encoder is used as a cross-modal teacher that encodes latent features as masked point modeling targets for 3D Transformer student representation learning.

Figure 1. We propose PartSLIP, a zero/few-shot method for 3D point cloud part segmentation by leveraging pretrained image-language models. The figure shows text prompts and corresponding semantic segmentation results (zoom in for details). Our method also supports part-level instance segmentation. See Figure 5 and Figure 7 for more results.

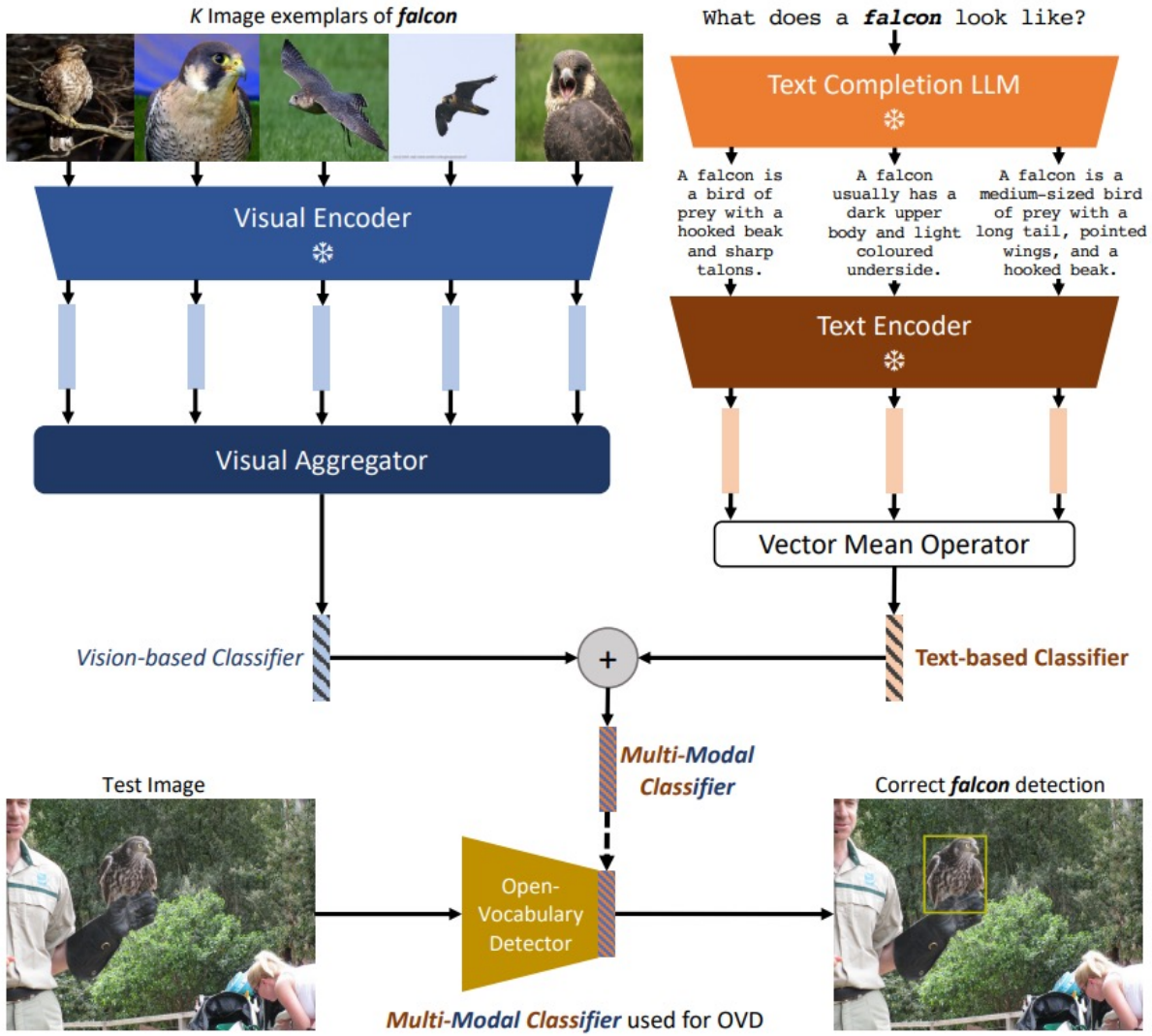


Figure 1. Overview of the architecture for generating text-based, vision-based, or multi-modal classifiers for OVOD. Vision (top left): A frozen visual encoder ingests image exemplars of **falcon** producing an embedding per exemplar. A trained aggregator takes these embeddings as input and produces a *vision-based classifier*. Text (top right): A text completion LLM is prompted to give descriptions of a **falcon** which are then encoded by a text encoder and averaged yielding a *text-based classifier*. Multi-Modal (middle): *Multi-Modal classifiers* are generated by adding the vision-based and text-based classifiers together. OVOD (bottom): The multi-modal classifier is used to detect the **falcon** in a standard model. Note, all three types of classifier: vision-based, text-based and multi-modal, can be used on the detector head for OVOD.

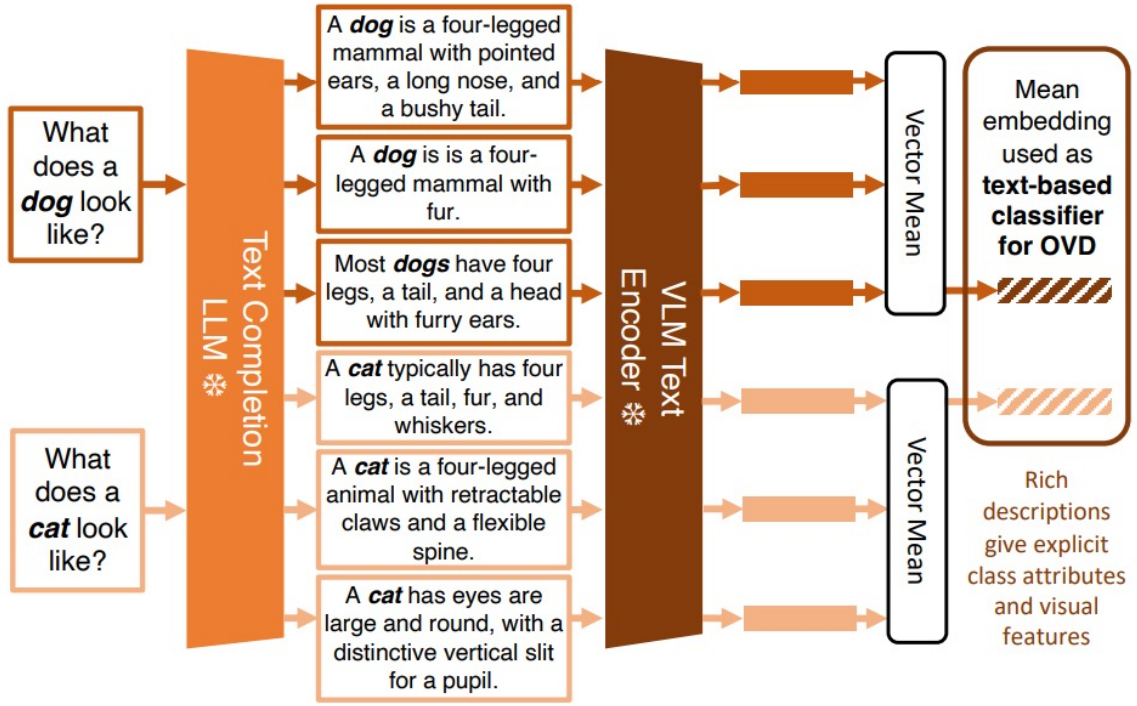


Figure 2. Generating powerful text-based classifiers. A LLM (GPT-3) is used to generate multiple rich descriptions of the class of interest. These descriptions are then encoded with the CLIP (Radford et al., 2021) VLM text encoder. The descriptions are more informative than the simple phrases, such as "(a photo of) a **dog**" or "(a photo of) a **cat**", used in previous work such as Detic and ViLD. Additional examples of class descriptions are given in the Appendix (Section F).

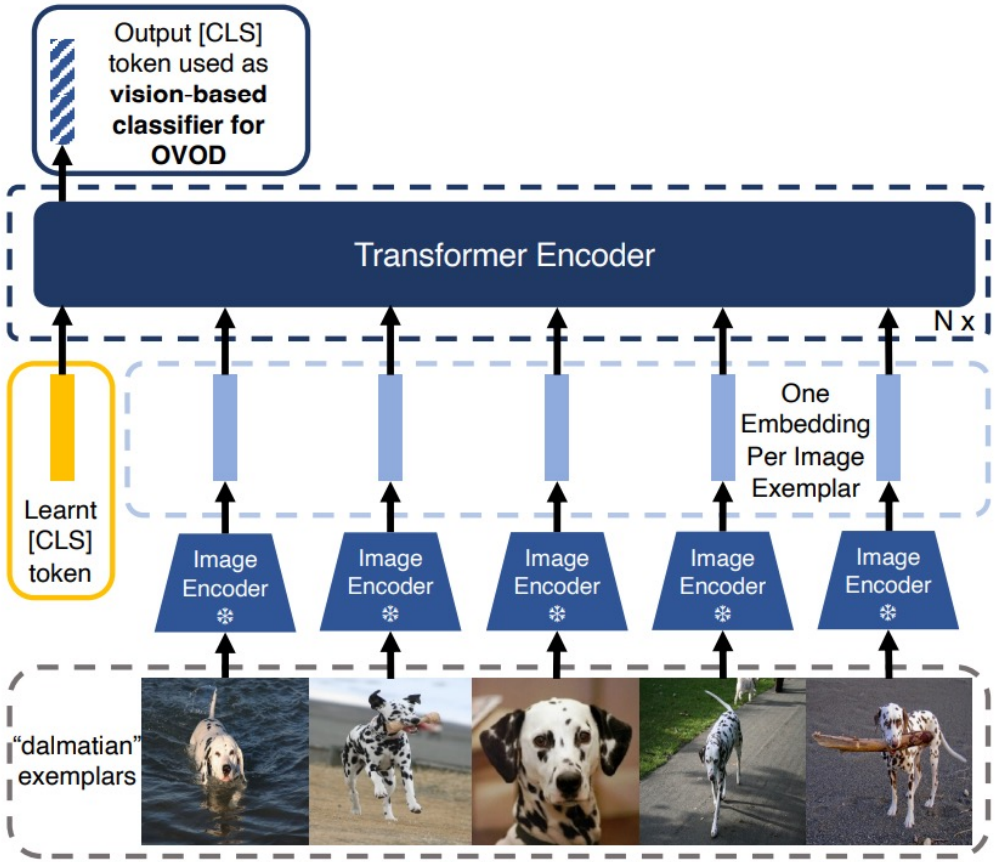


Figure 3. Generating an OVOD vision-based classifier from a set of image exemplars. A stack of transformer blocks is used to combine embeddings of multiple exemplars belonging to the same category.

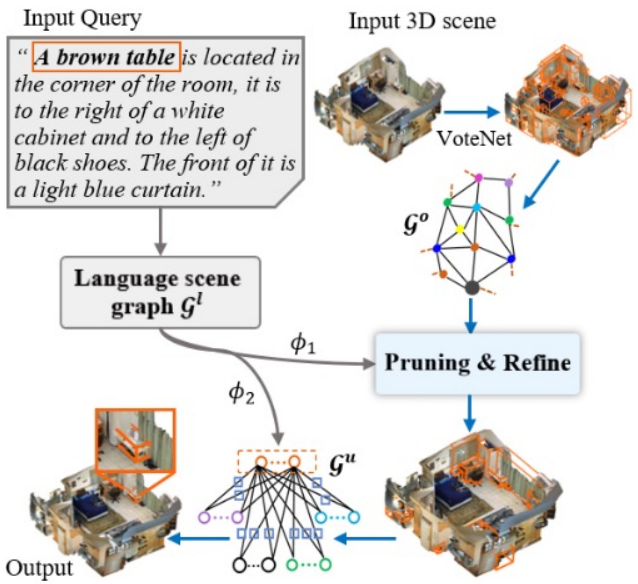


Figure 1. Proposed model for object grounding in 3D scenes. A multi-level proposal relation graph $\mathcal{G}^o$ is formed to strengthen the visual features of the initial proposals, then the 3D visual graph $\mathcal{G}^u$ is constructed under the guidance of the language scene graph $\mathcal{G}^l$ which refines the initial coarse proposals. The language scene graph $\mathcal{G}^l$ predicts the nodes matching with the 3D visual graph $\mathcal{G}^u$ and the matching scores ϕ_1 and ϕ_2 are fused to make the final grounding predictions.

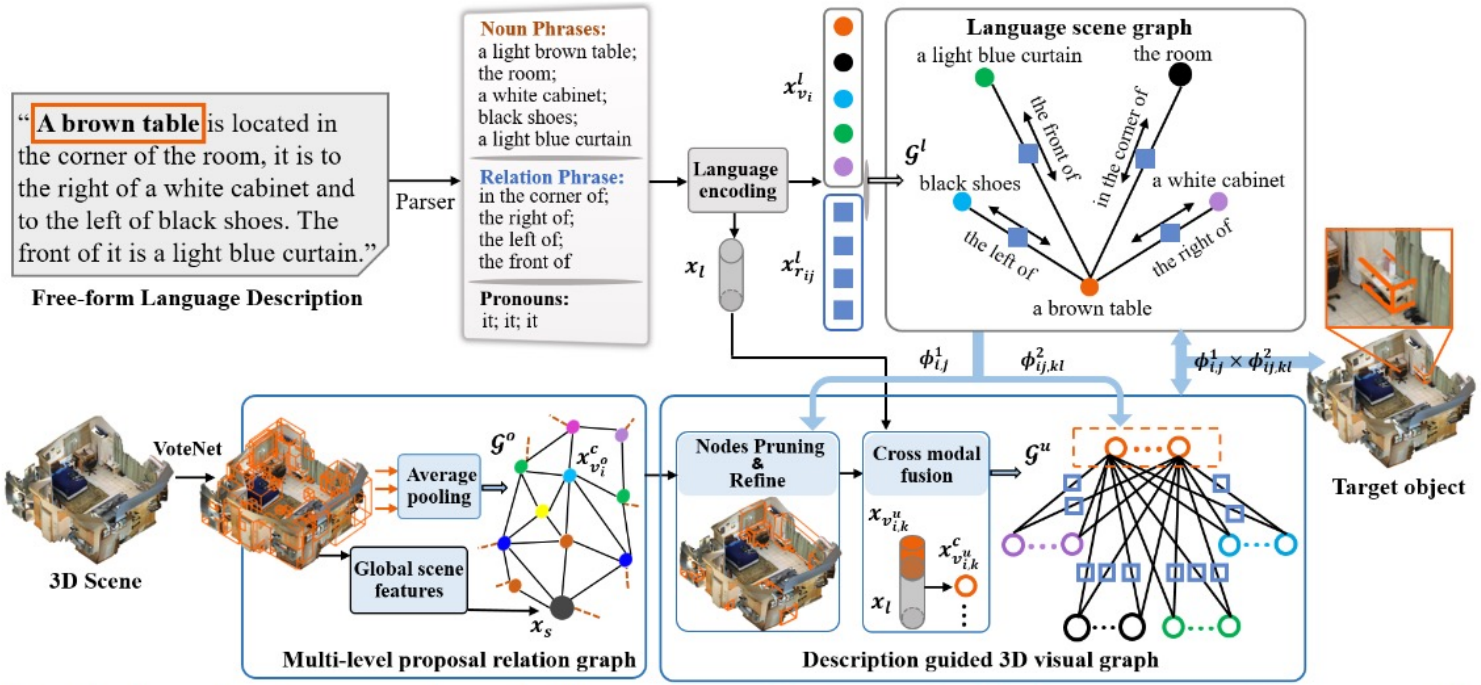


Figure 2. An overview of our proposed network. There are three modules in our method, the language scene graph $\mathcal{G}^l$ incorporates the rich structure and language context; the multi-context proposal relation graph leverages two occurrence relationships (object-object and object-scene) to strength the visual features of the initial proposals set; the description guided 3D visual graph $\mathcal{G}^u$ is defined on the pruned and refined proposals which is under the guidance of $\mathcal{G}^l$, then the nodes of $\mathcal{G}^u$ are adaptively matched with the nodes of $\mathcal{G}^l$, and then we fuse this with the matching score in proposals pruning for the final 3D object grounding.

- D3 Gallery: https://github.com/d3/d3/wiki/Gallery
- https://developers.google.com/chart/interactive/docs/gallery