# GlueGen: Plug and Play Multi-modal Encoders for X-to-image Generation

Can Qin[1,2], Ning Yu[1], Chen Xing[1], Shu Zhang[1], Zeyuan Chen[1], Stefano Ermon[3], Yun Fu[2], Caiming Xiong[1], Ran Xu[1]

[1]Salesforce Research [2]Northestern University [3]Stanford University

Code

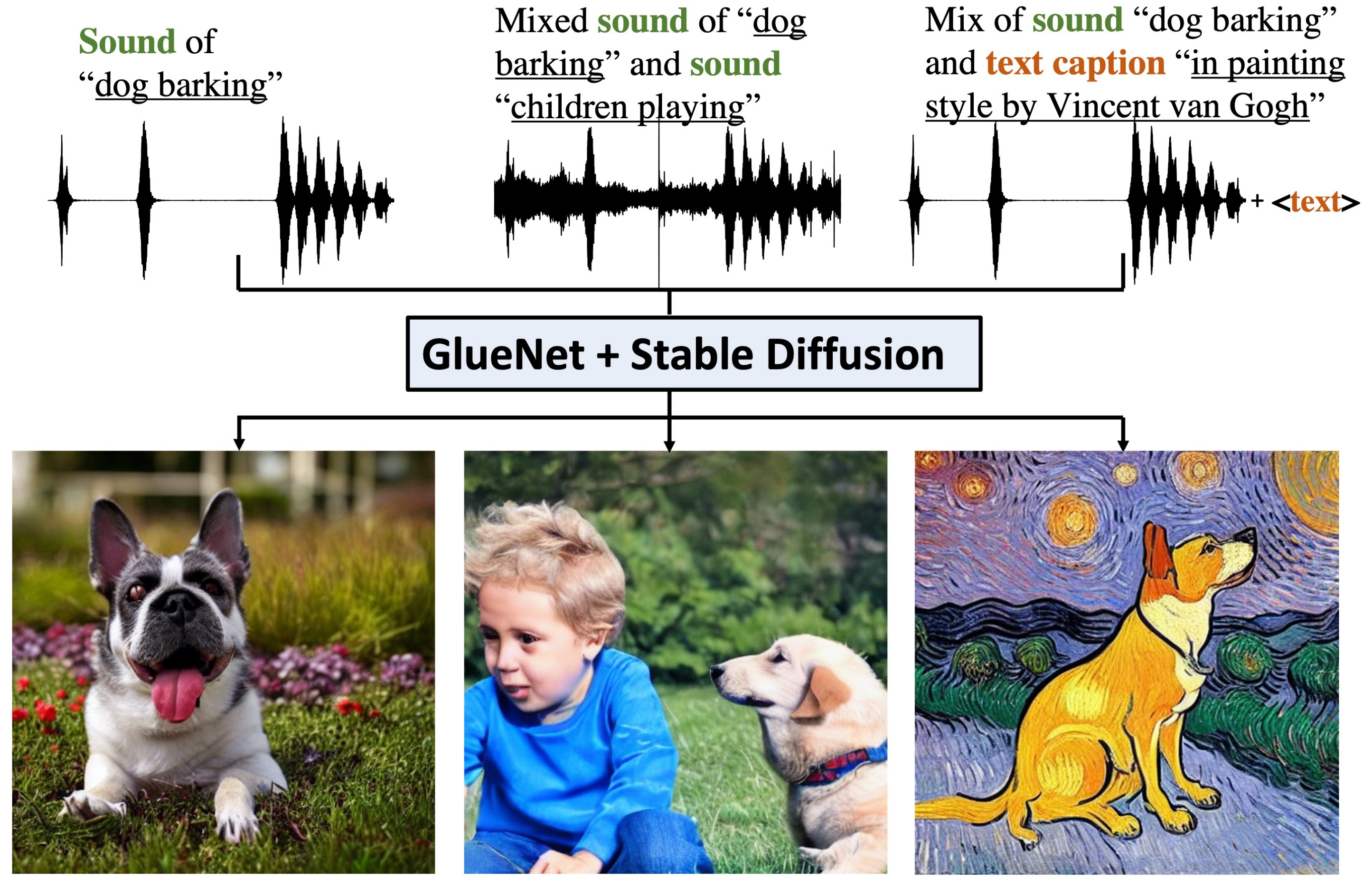
## Background

- Text-to-image (T2I) synthesis, generating photorealistic images from text prompts, has witnessed a tremendous surge in capabilities recently.

*"an astronaut riding a horse"*

## Motivation

- How to enhance the current text encoder of T2I model with more powerful language models?
- How to plug and play multi-modal encoders to enable X-to-image generation without time-consuming retraining?

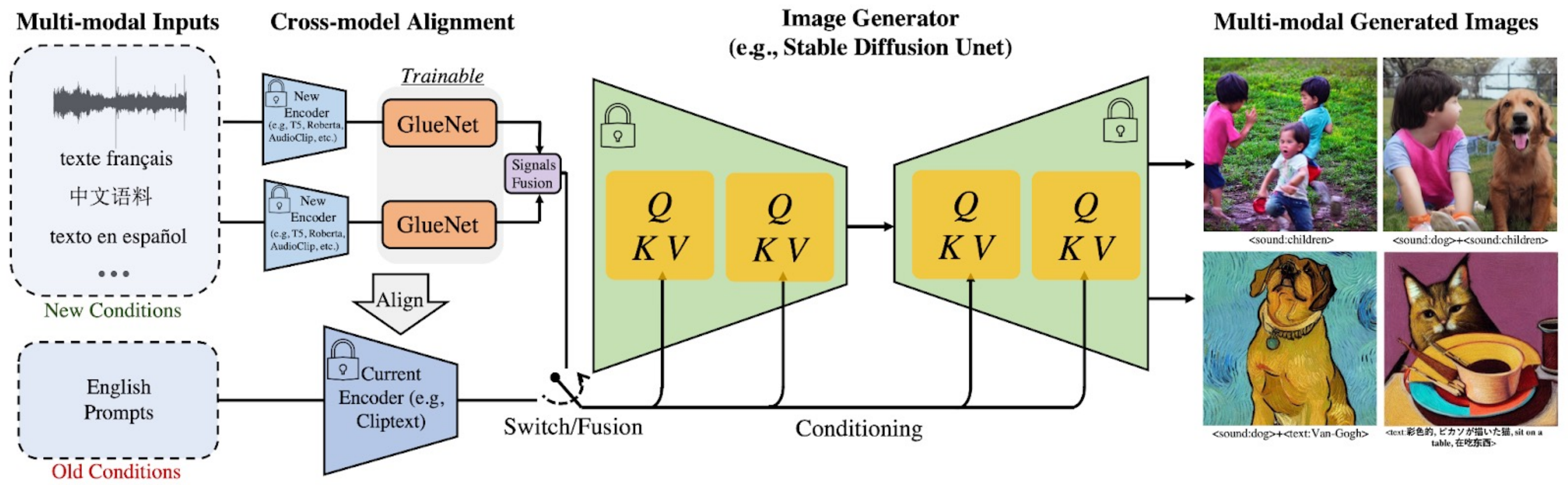

Sound of "dog barking"
Mixed sound of "dog barking" and sound "children playing"
Mix of sound "dog barking" and text caption "in painting style by Vincent van Gogh"
+ <text>
GlueNet + Stable Diffusion

(a) Single Sound (b) Mixed Sound (c) Sound-Text Mix

*Setting of GlueGen. GlueNet is trying to provide an adaptable portal for the Stable Diffusion model to input multi-modal data, such as text, audio, i.e., (a) and (b), or text-audio hybrid signals, i.e. (c), for X-to-image generation.*

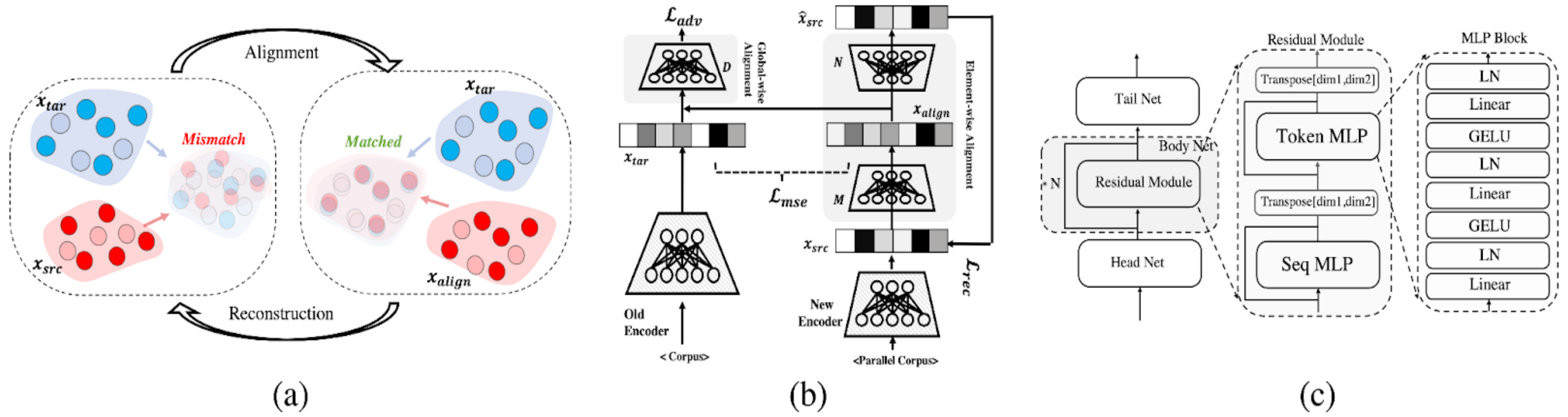
## Proposed Model

### Framework of GlueGen


Multi-modal Inputs
Cross-model Alignment
Image Generator (e.g., Stable Diffusion Unet)
Multi-modal Generated Images
texte français
中文语料
texto en español
New Conditions
New Encoder (e.g. T5, Roberta, AudioClip, etc.)
Trainable
GlueNet
GlueNet
Signals Fusion
Align
English Prompts
Old Conditions
Current Encoder (e.g, Cliptext)
Switch/Fusion
Conditioning
Q K V
<sound:children>
<sound:dog>+<sound:children>
<sound:dog>+<text:Van-Gogh>
<text:彩色的, ピカソが描いた猫, sit on a table, 在吃东西>

- **GlueGen** can plug in off-the-shelf pre-trained components, including:
  1. **More powerful language model:** T5-3B
  2. **Multi-lingual Language Models:** XLM-Roberta
  3. **Audio Encoders:** AudioCLIP

### Details of GlueNet



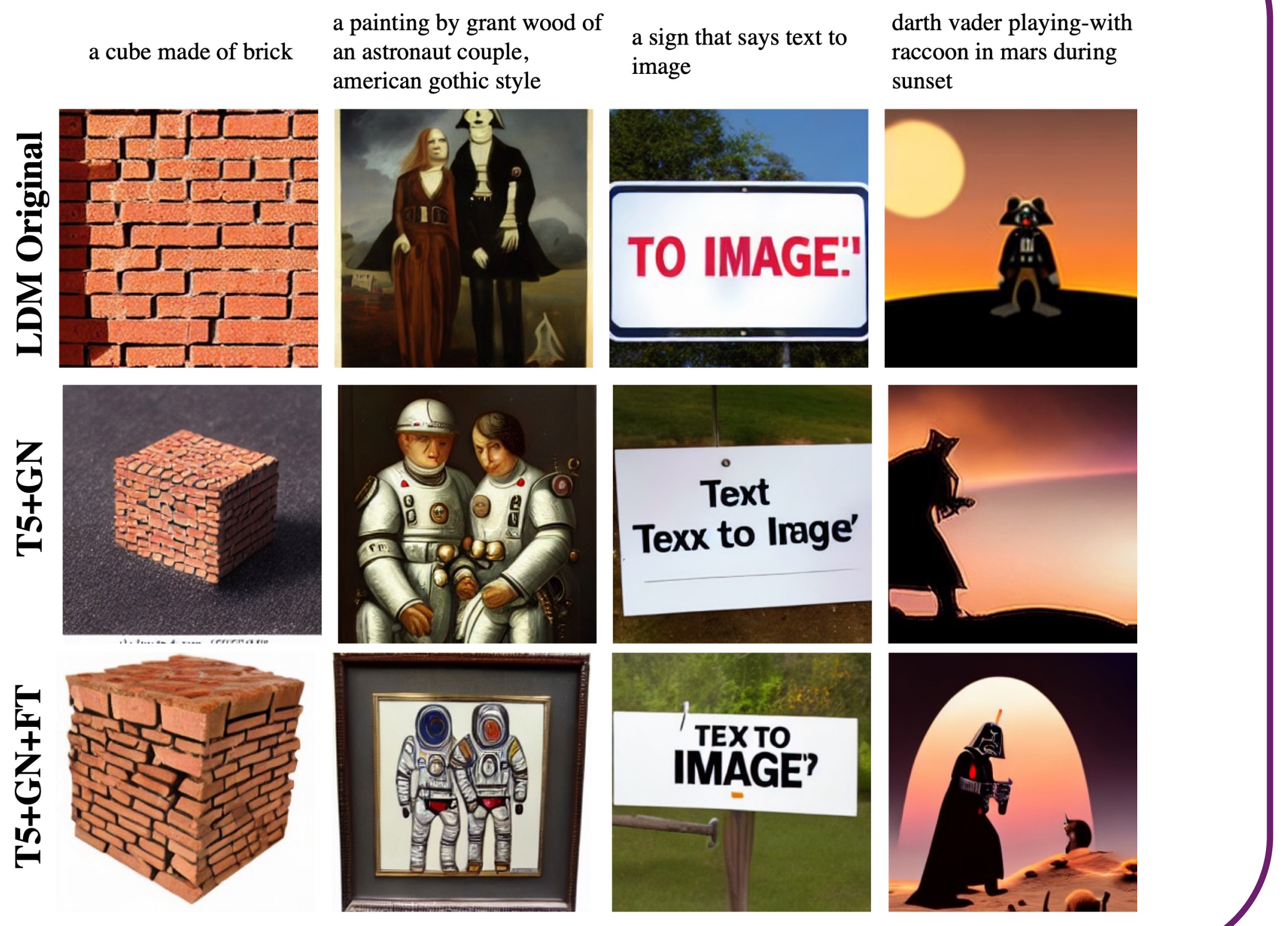
*(a) Illustration of features transformation throughout the model translation/alignment. (b) The general pipeline and learning objectives of our proposed GlueNet. (c) Detailed architecture of GlueNet Encoder/Decoder.*

- To achieve a desired **GlueNet**, we propose:
  1. **Encoder-decoder** structure as shown by (b)
  2. **Alignment and reconstruction** as illustrated by (a)
  3. **Dense regression module** with TokenMLP and SeqMLP as (c)
  4. **Without the need of re-training** Unet

## Monolingual Text-to-Image

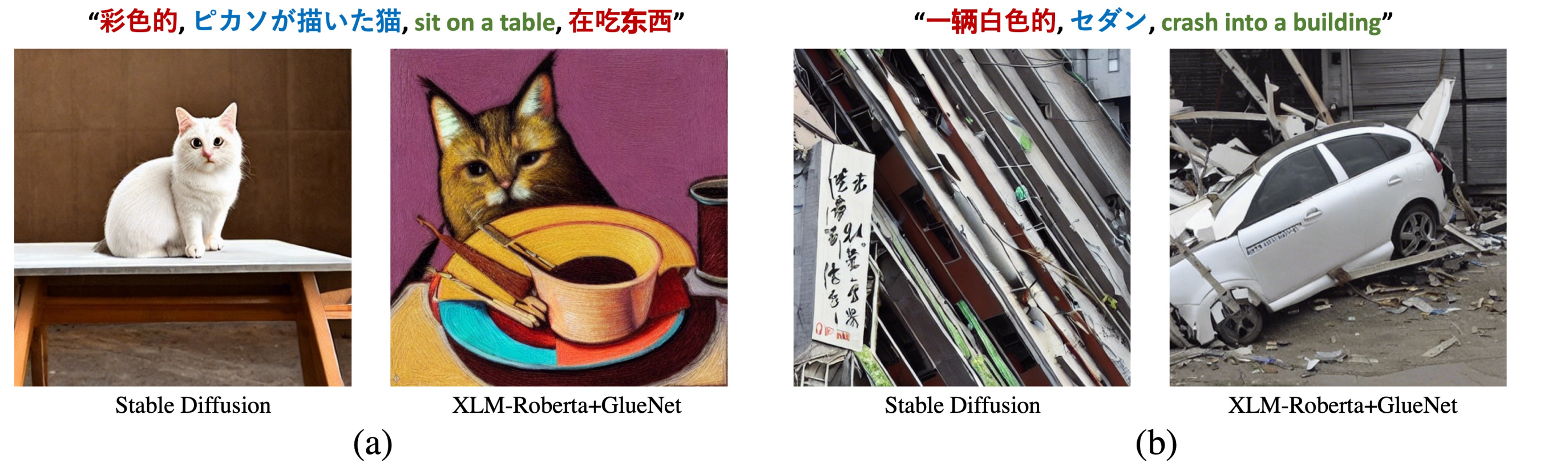
FID on COCO

| Method | FID↓ | ZS |
|---|---|---|
| CogView [13] | 27.10 | ✗ |
| LAFTTE [74] | 26.94 | ✗ |
| GLIDE [36] | 12.24 | ✗ |
| Make-A-Secne [16] | **11.84** | ✗ |
| LDM [45] | 12.63 | ✓ |
| LDM* | 13.55 | ✓ |
| T5+FT* | 23.30 | ✓ |
| T5+FT** | 12.41 | ✓ |
| T5+GlueNet | 14.32 | ✓ |
| T5+GlueNet+FT* | 12.05 | ✓ |


a cube made of brick
a painting by grant wood of an astronaut couple, american gothic style
a sign that says text to image
darth vader playing-with raccoon in mars during sunset
LDM Original
T5+GN
T5+GN+FT

## Multilingual Text-to-Image

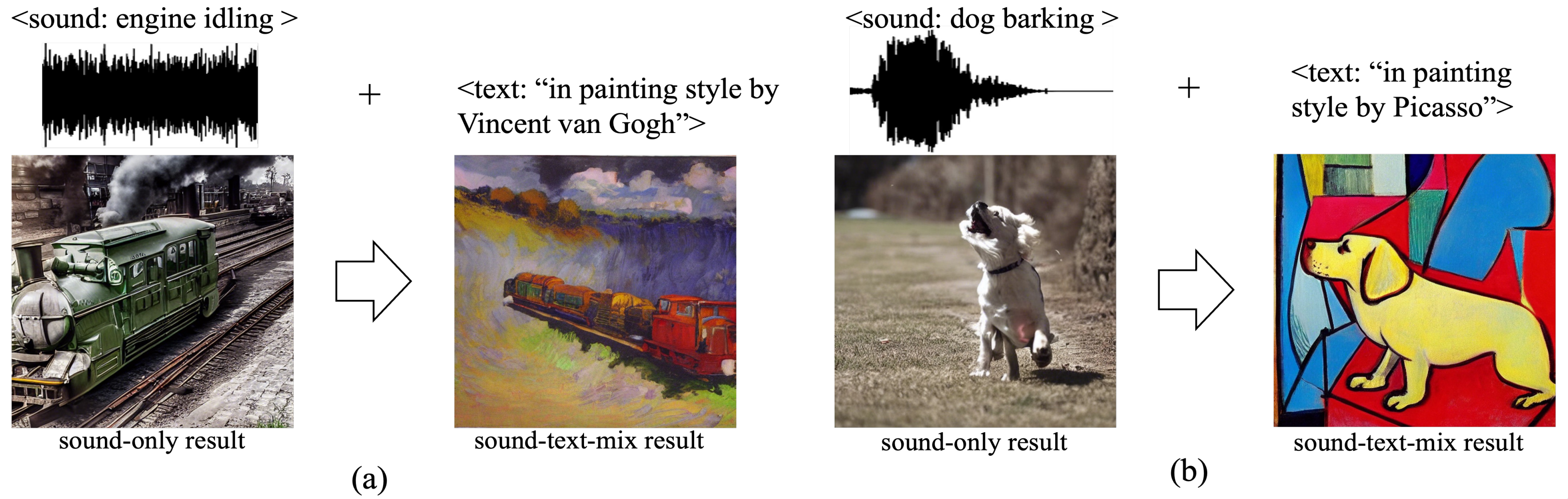
- Hybrid Multilingual Text-to-Image


"彩色的, ピカソが描いた猫, sit on a table, 在吃东西"
"一辆白色的, セダン, crash into a building"
Stable Diffusion
XLM-Roberta+GlueNet
(a)
(b)

- Multilingual Text-to-Image in Five Language Prompts

Chinese: 下午的花园的印象派绘画
French: Peinture impressionniste d'un jardin d'après-midi
Spanish: Pintura impresionista de un jardín de tarde
Japanese: 午後の庭の印象派絵画
Italian: Pittura impressionista di un giardino pomeridiano

## Sound and Sound-Text to Image


<sound: engine idling >
+
<text: "in painting style by Vincent van Gogh">
sound-only result
sound-text-mix result
(a)
<sound: dog barking >
<text: "in painting style by Picasso">
(b)