MaxFusion: Plug&Play Multi-Modal Generation in Text-to-Image Diffusion Models

Nithin Gopalakrishnan Nair†, Jeya Maria Jose Valanarasu*, Vishal M. Patel†

†Johns Hopkins University, *Stanford University

Motivation

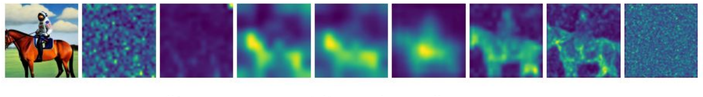
- Conditional generative models typically require large annotated training sets to achieve high-quality synthesis.
- Certain representations can be more effectively presented using simpler signals like pose maps, edges, or depth maps.
- The existing compositional algorithm is inefficient because it requires three forward passes through the diffusion models.
- We propose an efficient framework for spatial composition-based generation in diffusion models utilizing feature fusion.

Applications of MaxFusion

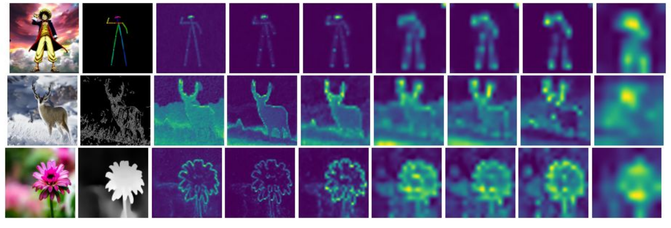
(a) Contradictory conditions {Task1 (Object 1) + Task2 (Object 2)} → Composite Task

Detecting Salient intermediate features in Diffusion UNet

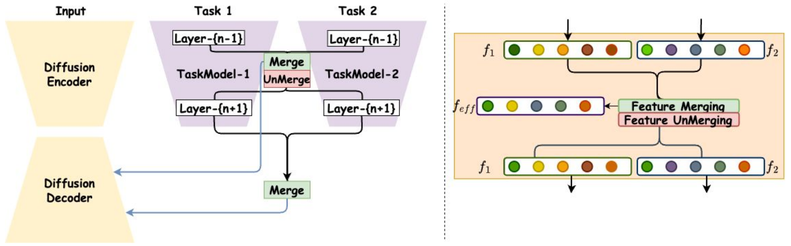
- Variance maps across the channels of a diffusion model capture the salient regions within the conditioning inputs.
- This conditioning method is effective for text prompts as well.



"An astronaut riding a horse"



Maxfusion



- The intermediate layer features of ControlNets for different tasks (such as depth and pose mapping) that are passed to the Diffusion U-Net are aligned.
- However, the individual layer outputs not passed to the main diffusion module may remain unaligned.
- We identify salient regions for merging based on feature variance maps.
- These regions are merged according to their relative importance and then passed to the Diffusion U-Net.
- The merged features are unmerged and passed to the individual ControlNets.

Algorithm

- Normalized channel variance based salient region estimation

$$\hat{\sigma}_i^{(j,k)} = \frac{\sigma_i^{(j,k)}}{\sum_{(j,k)} \sigma_i^{(j,k)}}$$

- Featu … sure:
- Interm … e merge intermediate features based on their correlation to integrate information from different modalities.

$$\rho^{(j,k)} = \frac{f_1^{(j,k)} \cdot f_2^{(j,k)}}{|f_1^{(j,k)}| \cdot |f_2^{(j,k)}|}$$

- **Variance based feature unmerging**: By unmerging features based on variance, we incorporate spatial regions that have higher conditioning intensity.

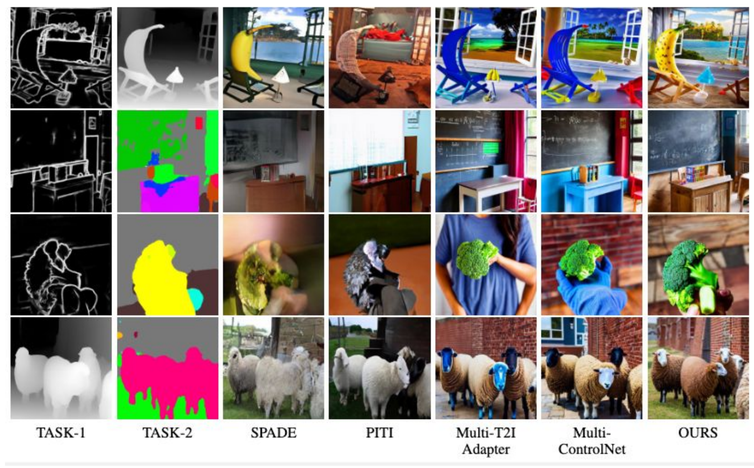
Algorithm 1 MaxFusion for scaling two modalities

Input: Model 1 layers l_1, Model 2 layers l_2 , threshold δ,

input condition c_1, input condition c_2
for $l = 1, \ldots, L$ **do**
  if $l = 1$ **then**
    $f_1 = l_1(c_1)$
    $f_2 = l_2(c_2)$
  else
    if $\rho(f_1^{(j,k)}, f_2^{(j,k)}) \geq \delta$ **then**
      $f_{eff}^{(j,k)} = \frac{f_1^{(j,k)} + f_2^{(j,k)}}{2}$
    else
      $f_{eff}^{(j,k)} = f_i^{(j,k)}, \max_i(\hat{\sigma}_i^{(j,k)})$
    end if
    SD decoder $\leftarrow feff$
    if $\rho(f_1^{(j,k)}, f_2^{(j,k)}) \geq \delta$ **then**
      $fi^{(j,k)} = f_{eff}^{(j,k)}; i = 1, 2$
    else
      $f_i^{(j,k)} = \begin{cases} f_i^{(j,k)} & i = \max_i(\hat{\sigma}_i^{(j,k)}) \\ \frac{\sigma_i^{(j,k)}}{\sigma_{max}^{(j,k)}} f_{max}^{(j,k)} & otherwise \end{cases}$
    end if
  end if
end for
return f_{eff}

Results

- Our algorithm enables seamless integration of different 2D spatial conditions to generate compositional cases.
- We consider two scenarios:
  - **Case 1**: The conditions do not agree with each other.
  - **Case 2**: The conditions agree with each other.



Ablations

- By varying the correlation threshold, we can adjust the influence of a condition at specific spatial locations.