MultiDiffusion: Fusing Diffusion Paths for Controlled Image Generation (ICML 2023)

Author: Omer Bar-Tal Lior Yariv Yaron Lipman Tali Dekel

Presenter: Yinghao Zhang, Keling Yao

Content

- Intro - collaborative painting
  - Can we do the same in image gen?
- Method - panorama
  - Merge patches
  - Math
  - Results

==========

- Improvement(patches -> region): region-based
  - Tight mask (bootstraping)
  - Results
- Combine: region-based + panorama
  - weakness
- Recap/takeaway

Prometheus Bound

🎨Frans Snyders

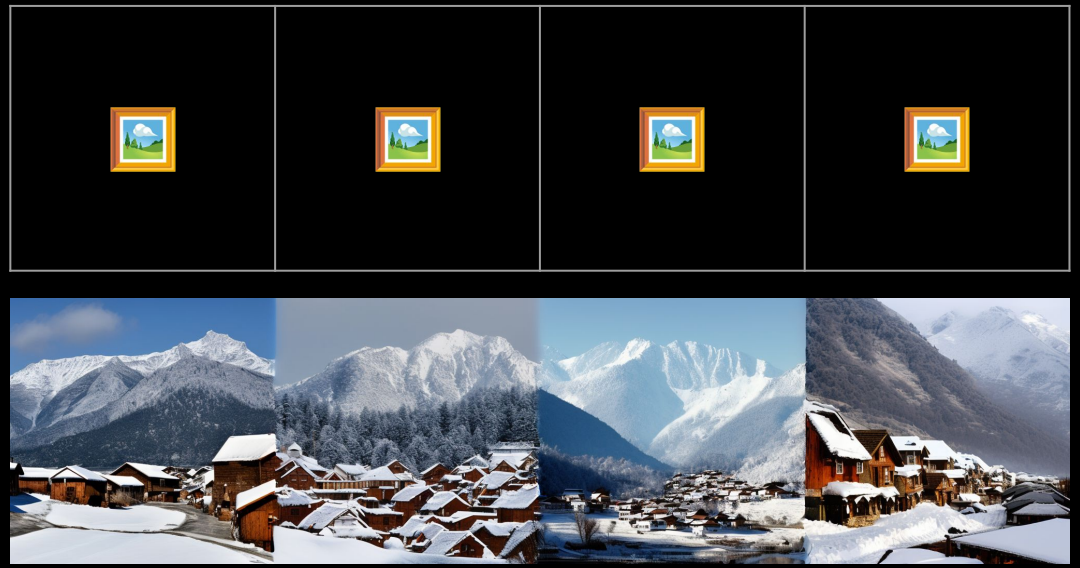
🎨Peter Paul Rubens

Source: Wikipedia - *Prometheus Bound* (Rubens)

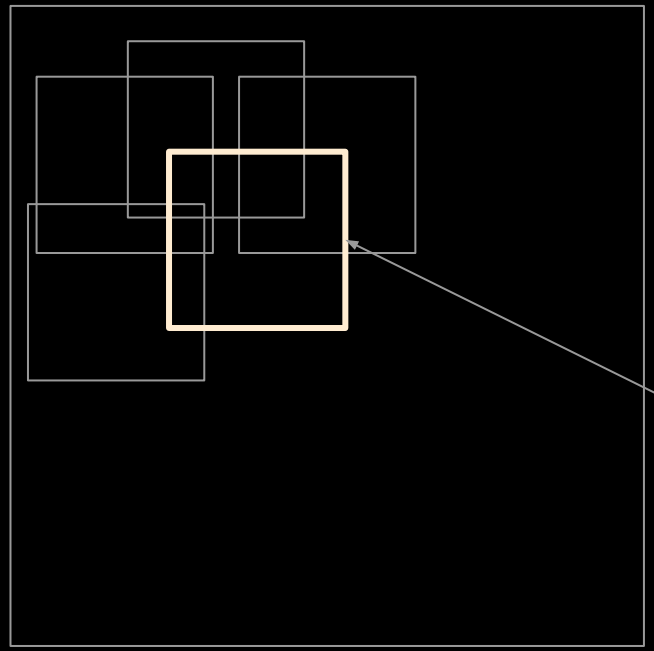
Collaborative image generation?

Goal: generate high-resolution images, using pre-trained low-resolution diffusion models, without any training

Idea 1:

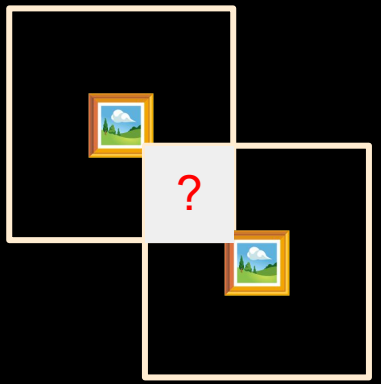


Idea 2: overlapping patches



Text input

Diffusion



Average?

Idea 2: not work…

Average

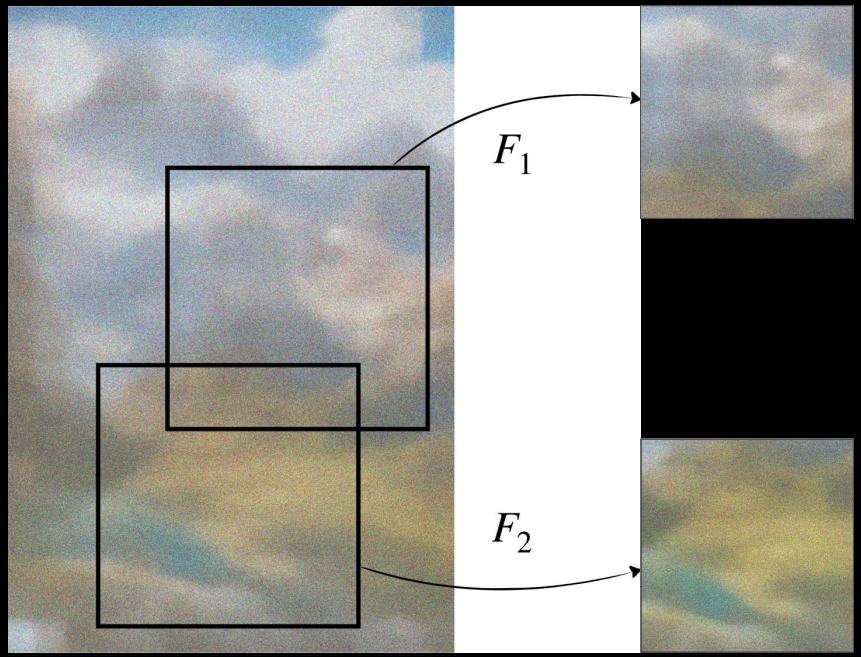
Idea 3: MultiDiffusion

❌ Generate patches, then average

✅ Average inside each denoising step

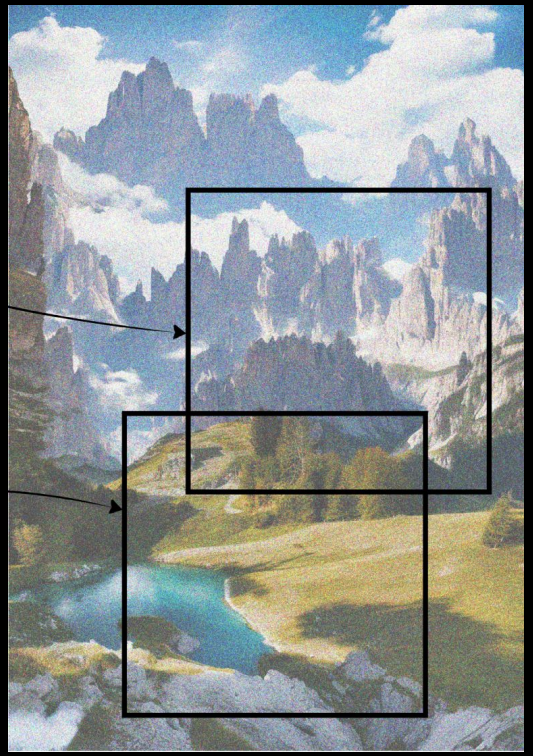
Idea 3: MultiDiffusion

In each denoising step…



Denoise

Denoise



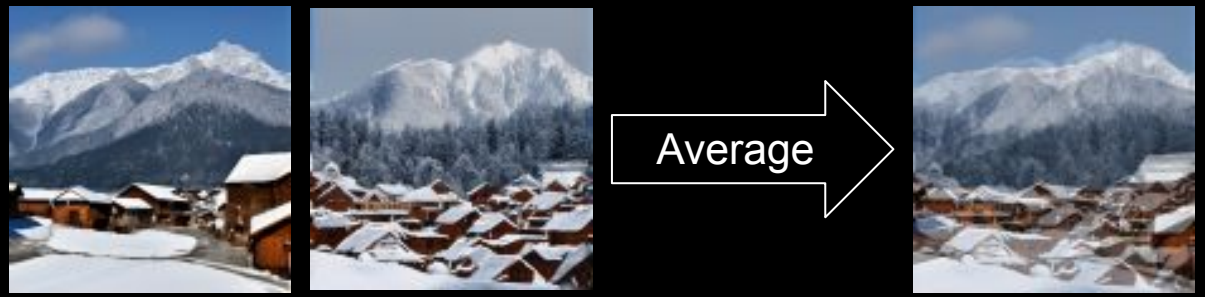
Idea 3: MultiDiffusion

512x2048

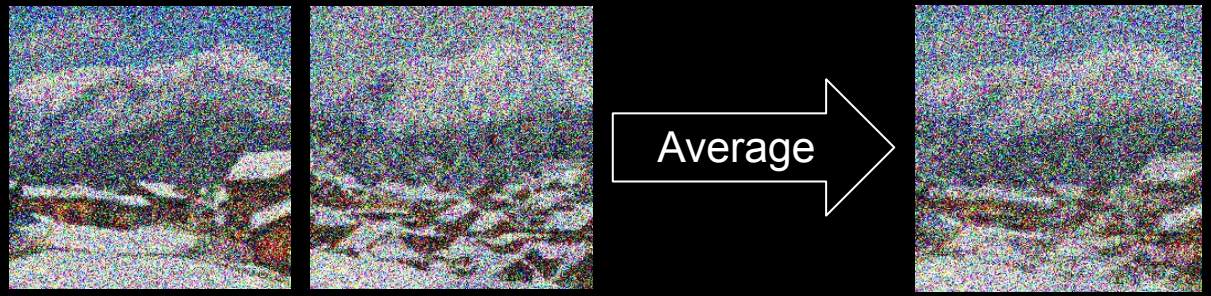
Backbone: Stable-Diffusion-2, 512x512

Analysis - why does this work?

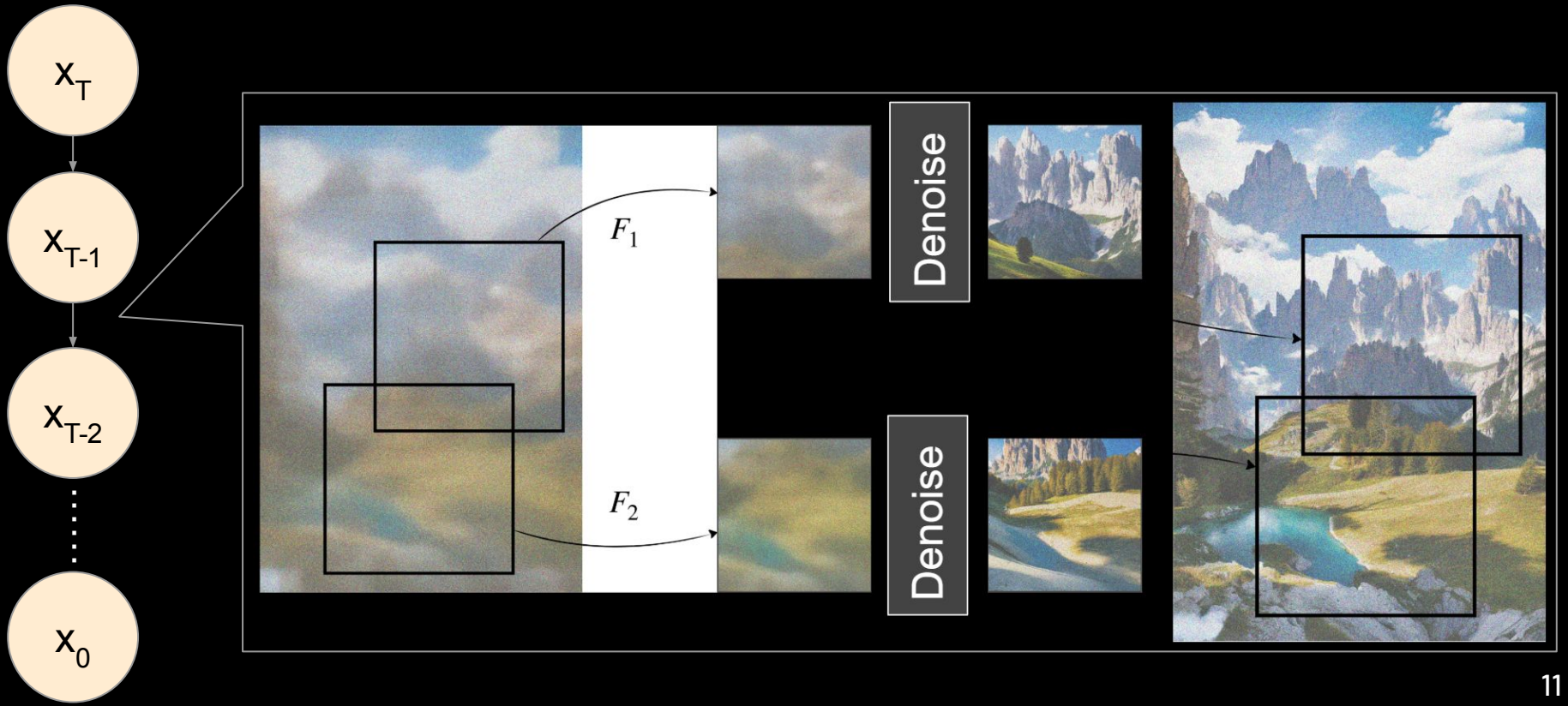
Idea 2 averaging final images:



Idea 3 averaging noised images:



Pipeline



Results

“a photo of the dolomites”

512x2048

Results

“a room full of cats”

512x2048

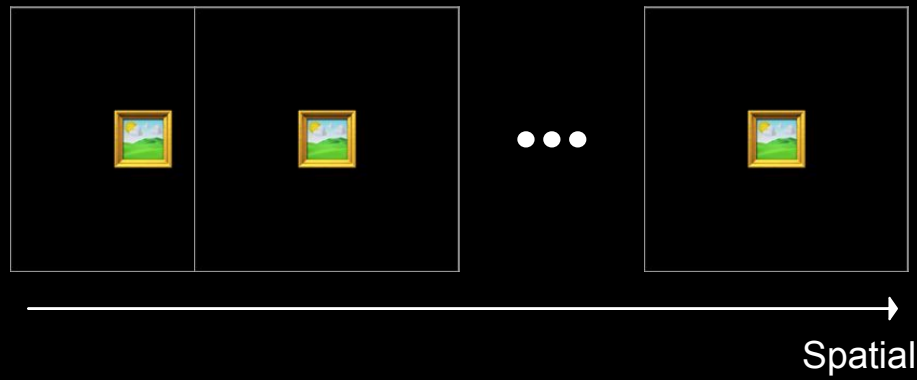
Results

“A giant glacier with dozens of penguins”

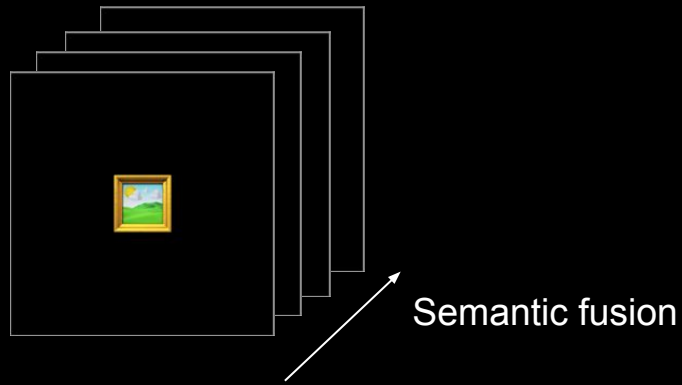
512x2048

Another Angle!!

Panorama:
Fuse Patches



Region based text2image:
Fuse semantic region

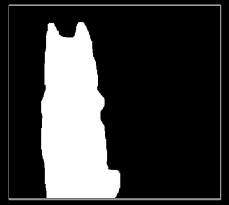


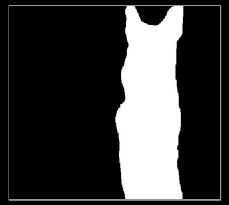
Region based text2image

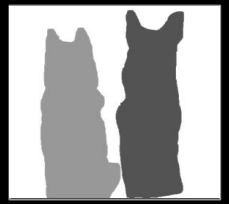
“A sunny day after the snow, two dogs sitting side by side, a German Shepherd dog on the right, a Husky dog on the left”

MultiDiffusion

Masks







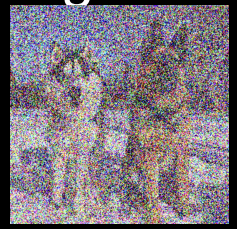
Prompt

“a Husky dog”

“a German Shepherd dog”

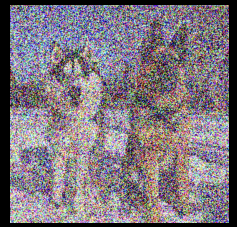
“A sunny day after snow”

Region based text2image



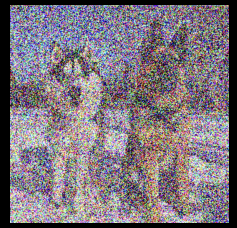
denoise →

“a Husky dog”



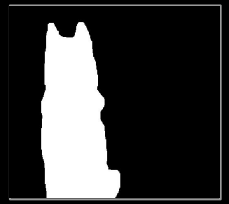
denoise →

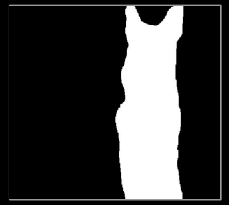
“a German Shepherd dog”

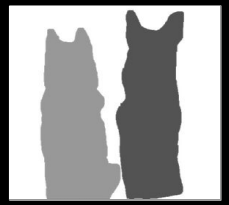


denoise →

“A sunny day after snow”

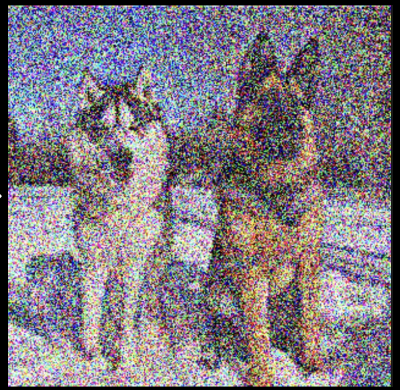






In each denoising step…

Weighted Average By masks



How to predict Semantic noises with strong masks boundary?

Problem: Post-denoising mask alignment

Bootstrapping: Fidelity to tight masks

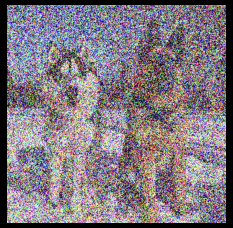
Bootstrapping: Fidelity to tight masks

Before T_{init} …

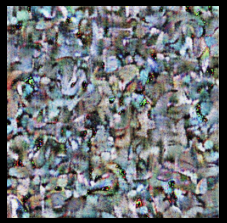
denoise

"a Husky dog"

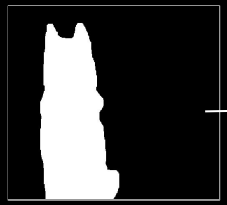
After T_{init} …



"a Husky dog"



Random background

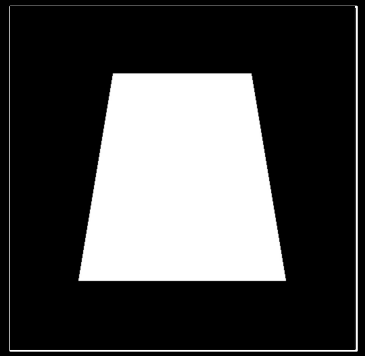


denoise

strong boundary!

Ablation study: Why Bootstrapping works?

Bg: “A stadium holding a soccer game, full of people.”



Fg Mask: “a green field”

$T_{init} = 0$

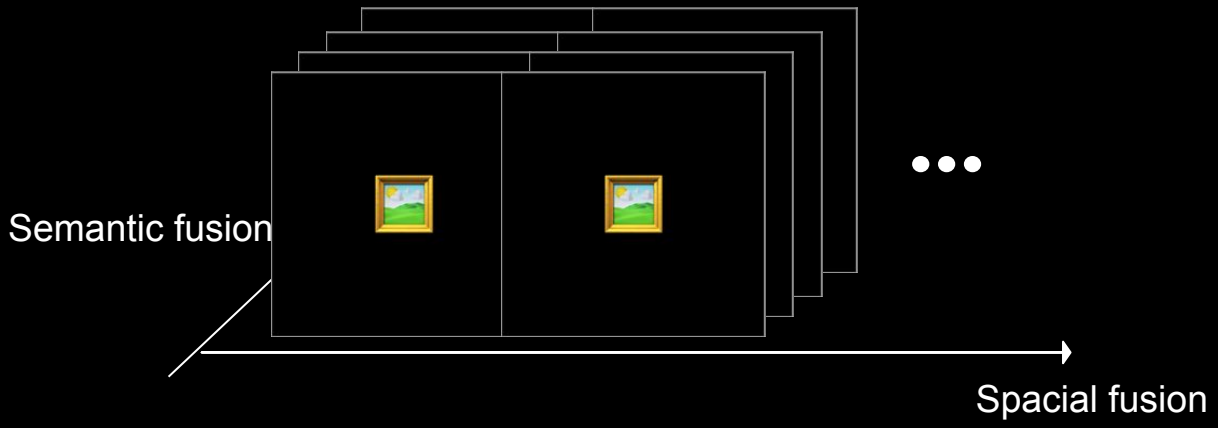
$T_{init} = 20$

$T_{init} = T_{total} = 50$

Tradeoff!
Global layout consistency
Local region accuracy

OUR experiment: region-based + panorama

Panorama with region control



Results (only Panorama)

“A stadium holding a soccer game, full of people.”

512x2048

Analysis

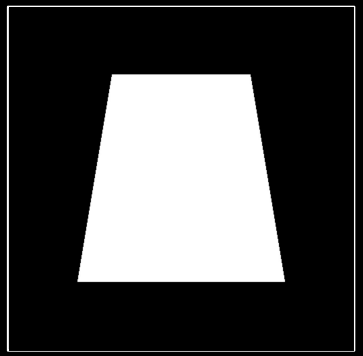
$\Phi(F_i(J\square), y)$

“A stadium holding a soccer game, full of people.”

All sliding windows conditioned on the same text!

Results (only Region based)

Bg: “A stadium holding a soccer game, full of people.”



Fg Mask: “a green field”

512x512

Results (region-based + panorama)

“A stadium holding a soccer game, full of people.”

Mask: “a green field”

512x2048

TakeAways

- ANY/High-Resolution image generation
- Training-Free paradigm
- Similar to ConvNet Sliding Windows (locality!)

Limitations

- Dependent on Prior Reference model
- Slow inference
- Semantic Consistency between different patches

Thank You!