VETIM: Expanding the Vocabulary of Text-to-Image Models only with Text

Martin Nicolas Everaert, Marco Bocchio, Sami Arpa, Sabine Süsstrunk, Radhakrishna Achanta

The problem we solve

Current methods [1, 2] cannot be used when sample images are not available. Instead, our method VETIM solely uses textual descriptions.

VETIM learns to **represent a complex concept as a single token** Sₙ. Optimisation with VETIM is **faster**.

VETIM **does not mimic** visual features from existing images.

Results

Our newly learned token Sₙ can be used instead of a lengthy text description t:

\[ t₁ = \text{a small, brilliant red stone that can produce the Elixir of Life and turn base metals into gold} \]

\[ t₂ = \text{a common, round fruit produced by the tree Malus domestica, cultivated in temperate climates} \]

\[ t₃ = \text{a twisted, abstract sculpture made of delicate, interlocking tendrils of glass} \]

Method

Our method VETIM optimises a **textual** Image Models only with Text encoder E.

\[ T[Sₙ] \rightarrow E(T[Sₙ]) \rightarrow \text{Image generation module (e.g. diffusion model)} \rightarrow \text{Generated image} \]

\[ \mathcal{L}_{\text{reg}} \]

\[ \mathcal{L}_{\text{sim}} \]

\[ \mathcal{L}_{\text{original}} \]

References

[3] Description generated with ChatGPT
[4] Definition of "apple" on wiktionary.org
[5] Description generated with ChatGPT

This work is supported by Innosuisse grant 48552.1 IP-ICT.

Project website: https://ivrl.github.io/vetim/