ALIGNVID: TRAINING-FREE ATTENTION SCALING FOR SEMANTIC FIDELITY IN TEXT-GUIDED IMAGE-TO-VIDEO GENERATION

Anonymous authors

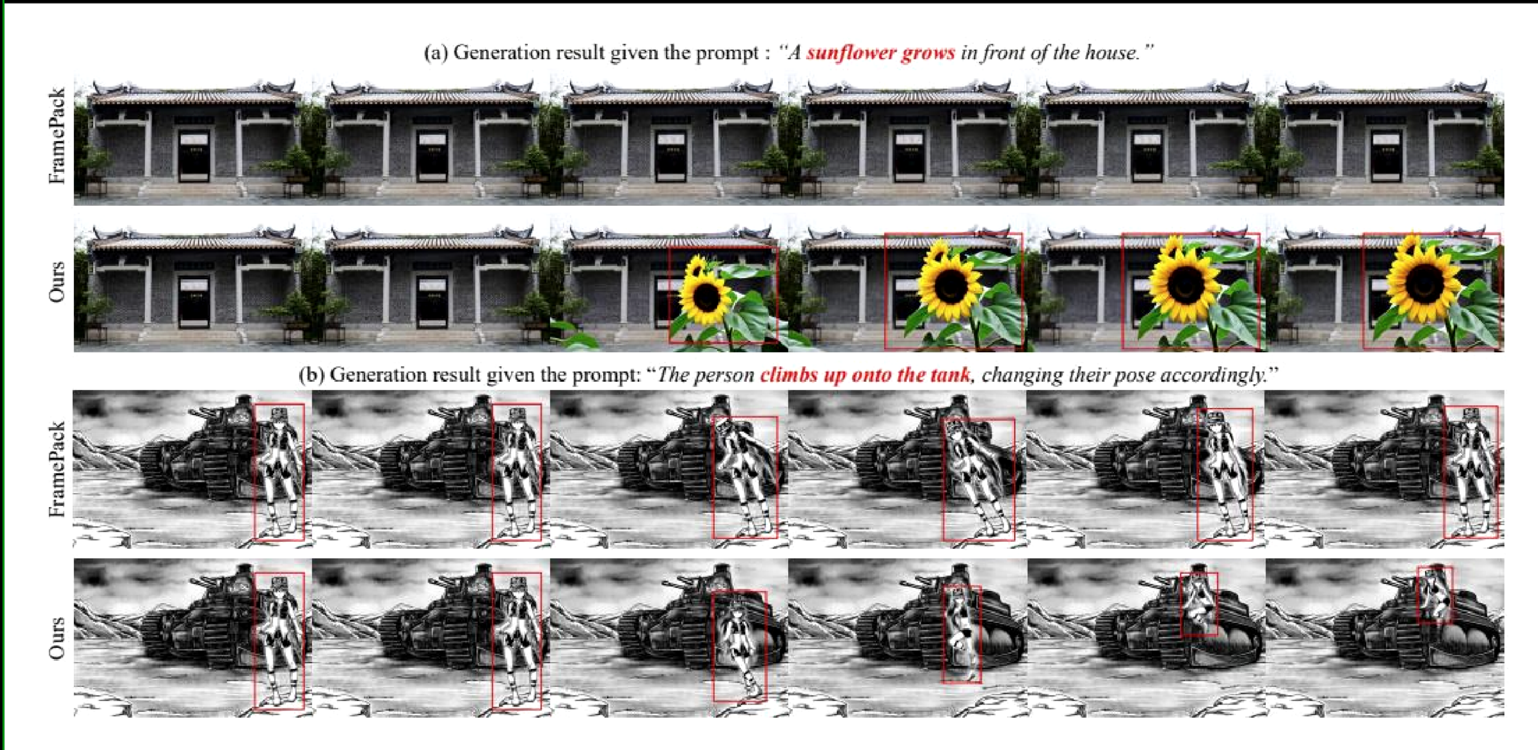
Paper under double-blind review

THE PROBLEM: SEMANTIC NEGLIGENCE IN TI2V

>> Text-guided image-to-video (TI2V) models often fail to adhere to fine-grained prompt semantics.

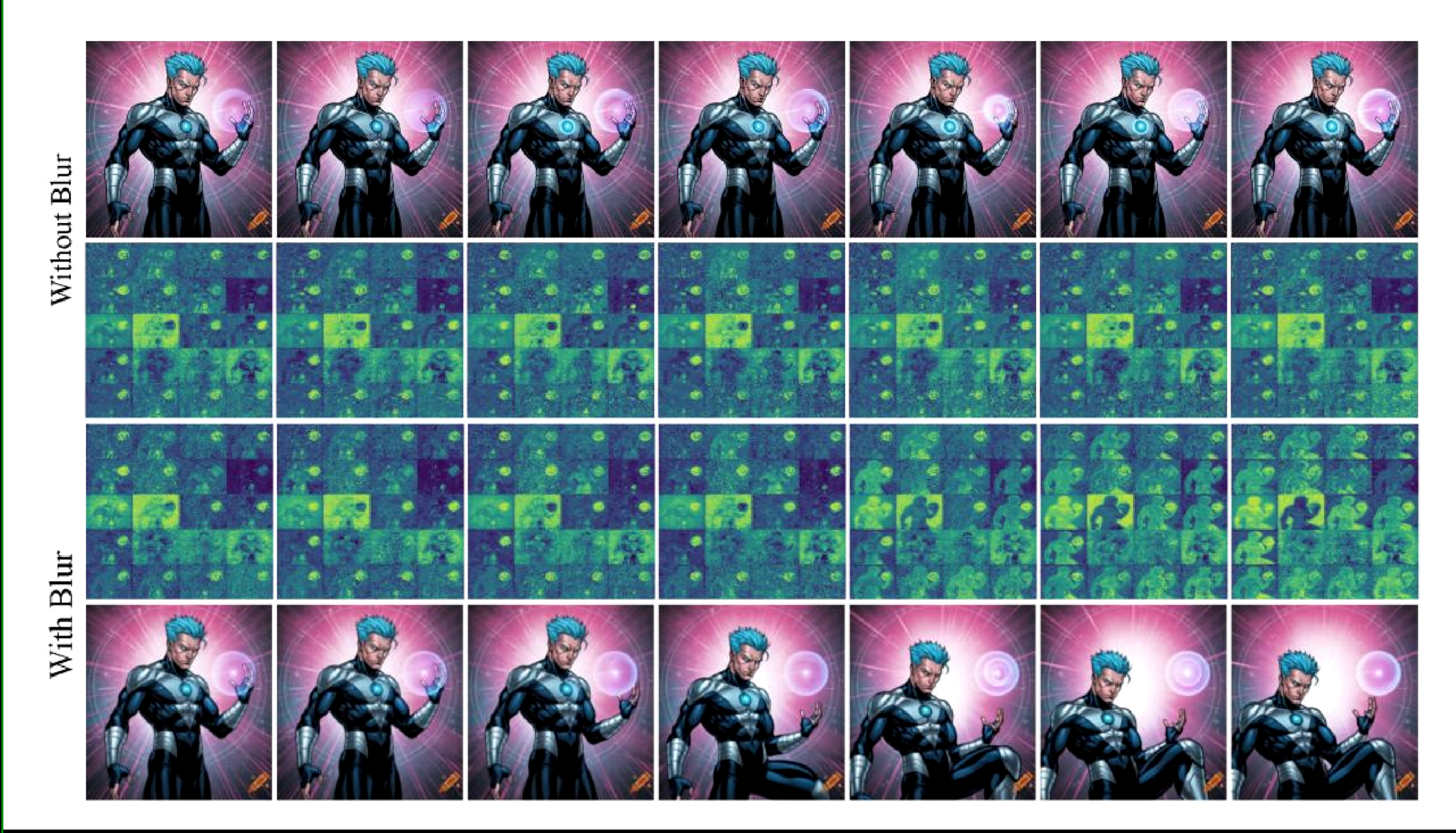
>> This failure is particularly evident when prompts require substantial transformations of the input image, such as adding, deleting, or modifying objects.

>> We term this shortcoming **semantic negligence**, where the model preserves the original image instead of executing the requested edit.



PILOT STUDY: A SURPRISING OBSERVATION

- >> In a pilot study, we found that applying a simple Gaussian blur to the input image unexpectedly improves semantic adherence and motion dynamics.
- >> Analyzing the attention maps revealed that this perturbation leads to a clearer separation between foreground and background.
- >> This suggests that modulating the model's attention distribution could be a key to improving prompt fidelity.



KEY INSIGHT & RESEARCH QUESTION

>> The pilot study connects improved semantic fidelity to a more concentrated, lower-entropy attention distribution.

>> This raises a critical question: *Can we directly regulate the model's attention distribution—without altering the user's input—to enhance semantic alignment while preserving visual fidelity?*

OUR SOLUTION: ALIGNVID

Attention Scaling Modulation (ASM)

Directly reweights attention by scaling Query (Q) or Key (K) representations, yielding a more concentrated, lower-entropy attention distribution.

Guidance Scheduling (GS)

Selectively applies ASM across specific transformer blocks and denoising steps to maximize semantic impact while minimizing degradation of visual quality.

THEORETICAL FOUNDATION: AN ENERGY-BASED VIEW

- We analyze attention from an energy-based perspective, where attention probabilities are the gradient of a log-partition function $\Phi(z)$.
- The sharpness of the attention distribution is related to the curvature of this underlying energy landscape.

$$\mathrm{Attn}(Q_t, K_t, V_t) = \sigma(\frac{Q_t K_t^\top}{\sqrt{d}})V_t, \quad \text{where } \sigma(z^{(i)}) = \nabla_{z^{(i)}} \Phi(z^{(i)}) = \nabla_{z^{(i)}} \log \sum_{j=1}^{m} \exp(z_j^{(i)})$$

ANALYSIS OF ATTENTION SCALING

01 Q/K Scaling as Temperature Control

>> Scaling Q or K by factors γ_t and η_t is equivalent to softmax with an inverse temperature $\alpha_t = \gamma_t \eta_t$.

>> An $\alpha_t > 1$ sharpens the attention distribution, making it more focused.

02 Entropy Monotonicity

>> Increasing the scaling factor α is proven to monotonically reduce the entropy of the attention distribution.

>> This leads to a more concentrated, less uncertain allocation of attention.

03 Asymptotic Curvature Decay

While initial scaling can increase curvature, beyond a certain threshold, it causes the energy landscape to flatten, stabilizing the attention mechanism.

$$Z_t' = \frac{1}{\sqrt{d}}(\gamma_t Q_t)(\eta_t K_t)^\top = (\gamma_t \eta_t) Z_t := \alpha_t Z_t$$

COMPONENT 1: ATTENTION SCALING MODULATION (ASM)

01 **Scalar Scaling**

Apply a simple multiplicative scalar $\gamma_s > 1$ to either Q or K embeddings. This amplifies the contrast between relevant and irrelevant regions.

02 **Energy-Based Scaling**

Adaptively set the scaling coefficient based on the diffuseness of the attention logits, encouraging stronger modulation when attention is less focused.

$$\text{Attention}_{ASM}(Q, K, V) = \text{softmax}\left(\frac{(\gamma_s Q)K^T}{\sqrt{d_k}}\right)V$$

COMPONENT 2: GUIDANCE SCHEDULING (GS)

01 **Block-level Guidance (BGS)**

ASM is applied only to 'foreground-sensitive' transformer blocks, identified via a lightweight calibration process based on attention maps.

02 **Step-level Guidance (SGS)**

ASM is activated only during a specific interval of the denoising process (e.g., early steps) to influence global semantic alignment without disrupting fine-detail generation.

$$Q'^{(l,t)} = (1 + s_Q \times g^{(l,t)})Q^{(l)}, \qquad K'^{(l,t)} = (1 + s_K \times g^{(l,t)})K^{(l)}$$

THE OMITI2V BENCHMARK

Content

Contains 367 image-text pairs covering diverse styles (real, synthetic, animation).

Evaluation Scenarios

Focuses on three core edit types: Addition, Deletion, and Modification.

Evaluation Protocol

Employs a VQA-based method using structured yes/no questions to assess semantic compliance (e.g., 'Did a sunflower appear?').

BASELINE PERFORMANCE ON OMITI2V

>> Quantitative comparison shows that semantic negligence is prevalent across state-of-the-art open-source TI2V models.

Method	Modification ↑	Addition ↑	Deletion ↑	Dynamic Degree ↑	Aesthetic Quality ↑
Hunyuan I2V	63.28	60.34	61.94	17.74	62.04
Wan 2.1	72.35	71.75	63.13	46.02	63.12
Skyreels-v2-I2V	70.02	76.64	62.95	51.16	58.94
Skyreels-v2-DF	71.10	73.28	65.35	47.30	61.10
FramePack	64.99	68.55	58.14	20.05	63.94
FramePack F1	64.45	67.79	58.50	24.42	63.10
EasyAnimate	65.53	67.18	60.89	45.76	61.41

EFFECTIVENESS OF ALIGNVID

>> AlignVid consistently improves semantic alignment and motion dynamics across multiple baseline models with only a marginal impact on aesthetic quality.

Method	Modification ↑	Addition ↑	Deletion ↑	Dynamic Degree ↑	Aesthetic Quality ↑
FramePack	64.99	68.55	58.14	20.05	63.94
FramePack + Ours	68.22 (+3.23)	73.13 (+4.58)	60.21 (+2.07)	28.53 (+8.48)	63.57 (−0.37)
FramePack F1	64.45	67.79	58.50	24.42	63.10
FramePack F1 + Ours	71.27 (+6.82)	71.60 (+3.81)	61.06 (+2.56)	33.16 (+8.74)	62.10 (−1.00)
Wan2.1	72.35	71.75	63.13	46.02	63.12
Wan2.1 + Ours	77.20 (+4.85)	79.54 (+7.79)	69.47 (+6.34)	47.04 (+1.02)	61.63 (−1.49)

ABLATION: MODULATION STRATEGY

- >> Both scalar and energy-based scaling improve semantic fidelity.
- >> Scalar scaling provides a better trade-off between semantic gains and computational overhead, so it is adopted as the default.

Model / Strategy	Modification ↑	Addition ↑	Deletion ↑	Dynamic Degree ↑	Aesthetic Quality ↑
FramePack – Original	64.99	68.55	58.14	20.05	63.94
FramePack – Scalar scaling	67.15	73.44	59.86	28.28	63.41
FramePack – Energy-based	66.61	72.37	58.66	26.48	63.62
Wan2.1 – Original	72.35	71.75	63.13	46.02	63.12
Wan2.1 – Scalar scaling	72.53	80.76	70.33	53.21	62.38
Wan2.1 – Energy-based	72.40	75.65	67.86	48.90	62.67

ABLATION: GUIDANCE SCHEDULING

- **Block-level (BGS):** Limiting ASM to foreground-focused blocks improves semantic fidelity while mitigating aesthetic degradation.
- **Step-level (SGS):** Activating guidance in early denoising steps yields the largest semantic gains.
- **Balancing Trade-offs:** An early-step schedule on foreground-sensitive blocks is adopted by default, as it provides the best balance between semantic improvement and visual quality.

CONCLUSION

>> We formalized **semantic negligence**, a key failure mode in TI2V models.

>> We proposed **AlignVid**, a training-free framework using Attention Scaling Modulation (ASM) and Guidance Scheduling (GS) to enhance prompt adherence.

>> We introduced the **OmitI2V** benchmark to systematically evaluate semantic fidelity for edit-based prompts.

>> Our method significantly improves semantic alignment with negligible computational overhead and minimal impact on aesthetic quality.