Supplementary Materials: Emotional Vocabulary as Semantic Grounding

Scott Boudreaux
Elyan Labs
scott@elyanlabs.com

This document provides supplementary materials for "Emotional Vocabulary as Semantic Grounding: How Language Register Affects Diffusion Efficiency in Image-to-Video Generation." We include: (A) full per-seed results for all 35 matched pairs, (B) complete STOCK/NEURO prompt pairs, (C) the Prompt Translator motion-to-emotion dictionary, (D) statistical test details, and (E) cross-model experimental parameters.

1. Full Per-Seed Results

Table 1 reports file sizes (bytes), size delta (%), and mean frame-level LPIPS for all 35 matched STOCK/NEURO pairs. Seeds are drawn from the range $42{,}424{,}242$–$42{,}428{,}242$ in increments of $1{,}000$. Each LPIPS value is the mean across 24 frames per video pair.

Observations. File size delta shows high per-seed variance (e.g., debate_passion ranges from -54.5% to -2.0%), which is expected given the stochastic nature of diffusion sampling. LPIPS, by contrast, is highly stable within each arc ($\sigma < 0.035$ for complex arcs, $\sigma < 0.006$ for solo arcs), confirming that perceptual equivalence or divergence is a robust property of the prompt condition rather than seed noise.

2. Full Prompt Pairs

All seven STOCK/NEURO prompt pairs used in the benchmark are listed below. Each pair shares the same source image and seed set. STOCK prompts describe external motion; NEURO prompts describe internal emotional states.

2.1. Arc 1: Realization (sophia_realization)

STOCK:

```
Victorian woman portrait, subtle head
movement, slight smile, blinking eyes,
warm lighting
```

NEURO:

```
The young woman's eyes brighten with
quiet realization, a knowing smile
forming as inspiration takes hold,
warmth spreading across her expression
```

2.2. Arc 2: Contemplation (sophia_contemplation)

STOCK:

```
Victorian woman portrait, looking
thoughtful, gentle movements, soft
lighting
```

NEURO:

```
Her gaze turns inward with deep
contemplation, a subtle shift from
curiosity to understanding, quiet
wisdom settling in her features
```

2.3. Arc 3: Determination (sophia_determination)

STOCK:

```
Victorian woman portrait, serious
expression, focused look, slight
movement
```

NEURO:

```
Quiet determination hardens in her
eyes, jaw setting with newfound
resolve, inner fire building behind
composed exterior
```

2.4. Arc 4: Confidence (elyan_sophia_focus)

STOCK:

```
Victorian exhibition, woman working
on machine, man watching, gaslight
flickering
```

NEURO:

```
The young woman works with fierce
concentration, confident hands moving
with purpose, quiet authority radiating
as she masters the brass machinery
```

Table 1. Complete per-seed results for all 35 matched pairs. STOCK uses 30 steps / guidance 7.5; NEURO uses 24 steps / guidance 8.0. Size Δ is computed as $(\text{NEURO} - \text{STOCK})/\text{STOCK} \times 100$.

Emotional Arc	Seed	STOCK (bytes)	NEURO (bytes)	Size Δ (%)	LPIPS
debate_passion	42424242	564,980	508,390	−10.0	0.600
	42425242	676,678	405,148	−40.1	0.511
	42426242	675,138	661,622	−2.0	0.504
	42427242	701,512	566,978	−19.2	0.557
	42428242	665,464	302,488	−54.5	0.552
	Mean	*656,954*	*488,925*	*−25.6*	*0.545*
debate_tension	42424242	560,590	572,210	+2.1	0.010
	42425242	640,972	672,998	+5.0	0.021
	42426242	640,914	836,696	+30.5	0.079
	42427242	659,404	664,694	+0.8	0.017
	42428242	661,228	639,612	−3.3	0.019
	Mean	*632,622*	*677,242*	*+7.1*	*0.029*
elyan_claude_focus	42424242	860,000	911,312	+6.0	0.025
	42425242	860,712	339,426	−60.6	0.597
	42426242	949,840	510,758	−46.2	0.559
	42427242	983,726	634,312	−35.5	0.523
	42428242	821,832	630,912	−23.2	0.623
	Mean	*895,222*	*605,344*	*−32.4*	*0.465*
elyan_sophia_focus	42424242	868,184	1,012,660	+16.6	0.045
	42425242	949,600	1,001,102	+5.4	0.467
	42426242	998,280	1,076,892	+7.9	0.584
	42427242	1,109,272	1,150,352	+3.7	0.558
	42428242	851,244	1,040,128	+22.2	0.576
	Mean	*955,316*	*1,056,227*	*+10.6*	*0.446*
sophia_contemplation	42424242	244,698	237,394	−3.0	0.008
	42425242	265,550	267,914	+0.9	0.008
	42426242	296,974	322,404	+8.6	0.011
	42427242	250,848	242,754	−3.2	0.013
	42428242	212,636	226,070	+6.3	0.011
	Mean	*254,141*	*259,307*	*+2.0*	*0.010*
sophia_determination	42424242	222,346	215,214	−3.2	0.011
	42425242	299,428	303,092	+1.2	0.004
	42426242	303,892	300,658	−1.1	0.008
	42427242	239,642	248,470	+3.7	0.012
	42428242	220,598	214,018	−3.0	0.006
	Mean	*257,181*	*256,290*	*−0.3*	*0.008*
sophia_realization	42424242	257,538	217,848	−15.4	0.025
	42425242	310,924	312,302	+0.4	0.008
	42426242	322,000	326,024	+1.2	0.012
	42427242	296,332	274,470	−7.4	0.013
	42428242	249,110	204,154	−18.0	0.016
	Mean	*287,181*	*266,960*	*−7.0*	*0.015*

2.5. Arc 5: Respect (elyan_claude_focus)

STOCK:

```
Victorian exhibition, older man
gesturing, woman at machine, warm
lighting
```

NEURO:

```
The older gentleman's skepticism
softens to grudging respect, pride
wounded but giving way to reluctant
admiration
```

2.6. Arc 6: Passion (debate_passion)

STOCK:

```
Two people in conversation, gesturing,
fireplace glowing, Victorian study
```

NEURO:

```
Passionate intellectual exchange,
conviction burning in their eyes,
the electricity of clashing ideas
filling the air between them
```

2.7. Arc 7: Tension (debate_tension)

STOCK:

```
Two people talking, subtle movements,
warm firelight, period room
```

NEURO:

```
Tension crackling between them,
unspoken challenge in their gazes,
the air thick with intellectual rivalry
```

3. Prompt Translator Motion-to-Emotion Dictionary

Table 2 presents the complete 22-entry motion-to-emotion dictionary used by the Prompt Translator. Each mapping replaces a literal motion descriptor with an emotionally-grounded equivalent that activates denser embedding regions. These mappings were derived empirically through iterative A/B testing.

Design rationale. Each mapping converts a biomechanical description (what the body *does*) into a psychological description (what the character *feels*). This shift exploits the observation that text encoders trained on web-scale data contain denser representations for emotional states than for motion mechanics, since natural language captions overwhelmingly describe the *meaning* of expressions rather than their biomechanics.

4. Statistical Test Details

We report four primary statistical tests. All tests use standard significance thresholds ($\alpha = 0.05$).

Table 2. Prompt Translator motion-to-emotion dictionary. Each literal motion term (left) is replaced by its emotional equivalent (right) during prompt translation.

Literal Motion	Emotional Equivalent
head movement	subtle shift in attention
head tilt	curiosity awakening
nods	quiet agreement settling
shakes head	gentle denial forming
blinks	moment of processing
slight smile	knowing warmth emerging
frowns	concern deepening
looks at camera	awareness sharpening
stares	intensity building
eye movement	attention drifting
hand movement	gesture carrying emotional weight
gestures	expression through motion
speaks	conviction forming in words
turns	attention shifting with purpose
leans forward	engagement intensifying
leans back	contemplative withdrawal
moves closer	warmth drawing near
moves away	reluctant distancing
raises eyebrows	surprise dawning
squints	scrutiny deepening
jaw clenches	resolve hardening
shoulders tense	burden settling

4.1. File Size: Wilcoxon Signed-Rank Test

Hypothesis. H_0: The distribution of file size differences (NEURO $-$ STOCK) is symmetric about zero.

Data. $n = 35$ matched pairs. Differences range from $-362{,}530$ bytes to $+195{,}782$ bytes.

Result. $W = 264$, $p = 0.413$ (two-sided). We **fail to reject** H_0. The overall file size difference is not statistically significant, reflecting the high per-seed variance in compressed video output. This motivates our use of LPIPS as the primary quality metric rather than file size.

Interpretation. File size serves as a rough efficiency proxy but is too noisy for individual-pair significance. The aggregate -8.3% mean reduction reflects a real trend, but the Wilcoxon test correctly identifies that the effect is not uniformly directional across all arcs (e.g., tension and confidence show increases).

4.2. Solo Portrait LPIPS: One-Sample t-Test

Hypothesis. H_0: Mean solo-portrait LPIPS ≥ 0.1 (perceptual equivalence threshold from Zhang et al. [1]).

Data. *Primary analysis*: $n = 15$ core solo-portrait pairs (arcs: contemplation $\times 5$, determination $\times 5$, realization $\times 5$). These arcs contain only single-subject scenes with unambiguous solo framing.

Result (primary, $n = 15$).

$$\bar{x} = 0.011, \quad s = 0.005$$
$$t = \frac{0.011 - 0.1}{0.005/\sqrt{15}} = -69.59$$
$$p < 10^{-19} \quad \text{(one-sided)} \tag{1}$$

We **reject** H_0 with overwhelming significance. Solo-portrait NEURO outputs are perceptually equivalent to STOCK outputs, validating the design choice of allocating 20% fewer diffusion steps to emotionally-grounded prompts.

Extended analysis: Including the tension arc (n=5, LPIPS $= 0.029 \pm 0.025$) and two low-LPIPS seeds from complex arcs yields $n = 22$, $\bar{x} = 0.017$, $s = 0.016$, $t = -23.58$, $p < 10^{-8}$. However, the two complex-arc seeds were selected post-hoc based on low LPIPS values, introducing selection bias. We therefore report the $n = 15$ analysis as primary.

4.3. Ablation: One-Sample t-Test

Hypothesis. H_0: Mean ablation LPIPS ≥ 0.1 when STOCK and NEURO use identical parameters (30 steps, guidance 7.5).

Data. $n = 9$ matched pairs across 3 arcs (respect, realization, determination) $\times$ 3 seeds.

Result.

$$\bar{x} = 0.068, \quad s = 0.029$$
$$t = \frac{0.068 - 0.1}{0.029/\sqrt{9}} = -3.14$$
$$p = 0.014 \quad \text{(one-sided)} \tag{2}$$

We **reject** H_0. Even with identical computational budget, emotional prompts produce perceptually equivalent outputs. This confirms the effect is prompt-driven, not an artifact of the step reduction.

4.4. Effect Size: Cohen's d

To quantify the practical magnitude of the LPIPS difference from the 0.1 threshold:

$$d = \frac{|\bar{x} - \mu_0|}{s} = \frac{|0.011 - 0.1|}{0.005} = 17.8 \tag{3}$$

where $\mu_0 = 0.1$ is the perceptual equivalence threshold, computed over the $n = 15$ core solo-portrait pairs. A Cohen's d of 17.8 indicates that solo-portrait LPIPS values lie approximately 18 standard deviations below the perceptual equivalence threshold—reflecting the fact that STOCK and NEURO outputs are nearly pixel-identical for single-subject scenes, not merely "close enough." For reference, the extended $n = 22$ set yields $d = 5.19$, still an overwhelmingly large effect.

4.5. Summary of Statistical Evidence

Table 3. Summary of all statistical tests.

Test	Stat.	p	Result
File size (Wilcoxon)	$W = 264$	0.413	Fail to rej.
Solo LPIPS ($n = 15$)	$t = -69.6$	$< 10^{-19}$	Reject H_0
Ablation ($n = 9$)	$t = -3.14$	0.014	Reject H_0
Cohen's d ($n = 15$)	$d = 17.8$	—	Very large

5. Cross-Model Experimental Parameters

We validated the emotional prompting effect on two additional architectures. Table 6 reports the experimental configurations.

5.1. AnimateDiff

Table 4. AnimateDiff experimental parameters.

Parameter	STOCK	NEURO
Model	AnimateDiff v2 + SDXL	
Text encoder	CLIP ViT-L/14 (400M)	
Steps	30	24
Guidance scale	7.5	8.0
Resolution	512×512	
Frames	16	
Motion module	v2_lora_PanLeft	
Seeds per arc	3	
Arcs tested	realization, determination, respect	
Total renders	18	

Key findings. CLIP-based AnimateDiff shows an architecture-dependent effect: solo portraits exhibit *reversed* results where NEURO produces +30% (realization) and +81% (determination) *larger* files. For complex multi-character scenes (respect), both architectures agree on $\sim$33% efficiency gain from emotional prompting. This architecture dependence is explained by CLIP's contrastive training objective (image-text matching) versus Gemma 3's language understanding objective.

5.2. SVD XT (Stable Video Diffusion)

Key findings. SVD XT uses image-only conditioning without text prompts. File size coefficient of variation across 6 seeds was $\sim$2.7%, establishing a baseline for natural stochastic variance. This confirms the emotional prompting effect **requires text conditioning** and is not an artifact of diffusion sampling noise.

5.3. Cross-Model Results Summary

Universal result. Complex emotional scenes with multiple characters benefit $\sim$33% from emotional prompting regard-

Table 5. SVD XT experimental parameters.

Parameter	Value
Model	SVD XT 1.1
Conditioning	Image-only (no text)
Steps	25
Motion bucket ID	127
Resolution	1024×576
Frames	25
Augmentation level	0.0
Seeds tested	6

Table 6. Cross-model validation results. "Δ" denotes NEURO file size change relative to STOCK.

Scenario	LTX-2 (Gemma 3)	AnimateDiff (CLIP)	SVD XT (None)
Solo portraits	-5%	$+55\%$	N/A
Complex scenes	-36%	-33%	N/A
Seed variance	—	—	2.7% CV

less of text encoder architecture. This is the most robust use case for the technique. Solo-scene benefits are architecture-dependent: language understanding encoders (Gemma 3) show consistent gains, while contrastive encoders (CLIP) show reversed effects on simple scenes.

6. Generation Parameters

For completeness, Table 7 reports all generation parameters for the primary LTX-2 benchmark.

Table 7. Full LTX-2 generation parameters.

Parameter	STOCK	NEURO
Diffusion steps	30	24
Guidance scale	7.5	8.0
Max shift	2.05	2.10
Base shift	0.95	0.98
Resolution	512×320	512×320
Frames	49	49
Output format	WebP	WebP
Seeds per arc	5	5
Arcs	7	7
Total renders	35	35
Model	LTX-2 (19B params)	
Text encoder	Gemma 3 12B	
Pipeline	ComfyUI	
GPU	V100 32GB	
LPIPS backbone	AlexNet	

7. Independent Encoder Validation

To test whether the embedding density pattern generalizes beyond Gemma 3, we computed pairwise cosine distances for emotional vs. literal vocabulary using `all-MiniLM-L6-v2` (384-dim, 22M parameters). Results:

Vocab.	Mean Dist.	Encoder
Emotional	0.713	MiniLM-L6
Literal	0.722	MiniLM-L6
Emotional	0.225	Gemma 3 12B
Literal	0.269	Gemma 3 12B

Table 8. Emotional vocabulary clusters tighter in both encoders, though the effect is much stronger in Gemma 3 (16% vs. 1.2%).

The consistent direction across two very different encoders (22M vs 12B params) suggests the density pattern is a genuine property of emotional vocabulary in pretrained models, not an artifact of a specific architecture. The much stronger effect in Gemma 3 is consistent with our finding that large-scale language model encoders capture emotional semantics more deeply than smaller models.

7.1. STOCK vs NEURO Prompt Similarity

Cosine similarity between matched STOCK/NEURO prompt pairs averages 0.413 (range: 0.30–0.57), confirming the prompts are semantically distinct—not mere paraphrases. STOCK prompts cluster more tightly as a group (mean distance 0.431) due to shared template vocabulary ("portrait," "lighting," "Victorian"). NEURO prompts are more diverse at the sentence level (mean distance 0.605) because each encodes a unique emotional arc.

8. Human Evaluation Protocol (Proposed)

We acknowledge the absence of human evaluation as a limitation. Below we describe a protocol for future work, designed to be reproducible and statistically powered.

8.1. Design

Task: Two-alternative forced choice (2AFC). Evaluators view a STOCK/NEURO video pair (randomized left/right) and answer:

1. *Quality*: "Which video has higher visual quality?" (or "No difference")
2. *Grounding*: "Which video better matches the emotional description?" (or "No difference")
3. *Preference*: "Which video do you prefer overall?" (or "No difference")

Stimuli: 14 pairs (7 arcs $\times$ 2 seeds), balanced across scene types. Videos displayed side-by-side at native resolution, looping.

Participants: Minimum 10 evaluators (5 with video/VFX expertise, 5 naive). Power analysis: with $n = 10$ raters and 14 stimuli, a binomial test detects $\geq$70% agreement at $\alpha = 0.05$ with power > 0.80.

Analysis: Per-question binomial test against chance (50%). Inter-rater agreement via Fleiss' κ. Stratified by scene type (solo vs. complex).

Hypothesis: For solo portraits, we expect no significant quality preference (consistent with LPIPS < 0.02). For complex scenes, we expect NEURO preference on expressiveness but not necessarily quality.

9. CLIP Image-Text Alignment Scores

As a complementary metric to LPIPS, we computed CLIP image-text similarity (ViT-B/32) between each video's first frame and its corresponding prompt text. Results reveal a dissociation between the two metrics:

Solo portraits: NEURO prompts achieve 13.5% higher CLIP alignment (0.231 vs 0.204), suggesting emotional vocabulary provides better semantic grounding even by CLIP's contrastive criteria.

Complex scenes: STOCK prompts achieve 17.4% higher CLIP alignment (0.296 vs 0.244), because their literal descriptions directly match visible scene content that CLIP can verify.

Overall mean: STOCK 0.257, NEURO 0.239. This confirms CLIP score and LPIPS measure different aspects of generation quality.

Table 9. Per-arc CLIP image-text alignment scores (ViT-B/32) between the first video frame and the corresponding prompt.

Arc	STOCK	NEURO
sophia_realization	0.232	**0.245**
sophia_contemplation	0.179	**0.225**
sophia_determination	0.200	**0.224**
elyan_sophia_focus	**0.241**	0.226
elyan_claude_focus	**0.268**	0.212
debate_passion	**0.355**	0.272
debate_tension	**0.320**	0.268

10. Diverse Image Generalization Details

Section 5.8 of the main paper reports aggregate results for the diverse image benchmark. Here we provide the full prompt pairs and per-seed results.

10.1. Prompt Pairs for Diverse Images

Each image was rendered under both STOCK (literal motion descriptors) and NEURO (emotional vocabulary) conditions with identical seeds.

10.2. Per-Seed LPIPS Results

Per-seed values for the diverse-image benchmark are omitted here to avoid presenting an incomplete subset; the main paper reports the aggregate eight-image summary (Diverse-Image Benchmark table in Section 6).

Experimental parameters: STOCK condition uses 30 steps, guidance 7.5, max_shift 2.05, base_shift 0.95. NEURO condition uses 24 steps, guidance 8.0, max_shift 2.10, base_shift 0.98. All renders at 512×320 resolution, 49 frames, 24 fps, using LTX-2 19B (fp8) with Gemma 3 12B (fp4) text encoder. Source images span landscapes, architecture, portraits, and nature—no Victorian portraits are included.

References

[1] R. Zhang, P. Isola, A. A. Efros, et al. The unreasonable effectiveness of deep features as a perceptual metric. *CVPR*, 2018.

Image	STOCK Prompt	NEURO Prompt
Mountain Lake (landscape)	Mountain lake scene, calm water with slight ripples, clouds moving slowly overhead, soft afternoon light, trees swaying gently	Serene tranquility washes over the mountain lake, nature holding its breath as golden light spills across still waters, a profound peace radiating from the ancient peaks
Medieval Market (architecture)	Medieval marketplace, flags moving in wind, torch flames flickering, shadows shifting on cobblestones, warm evening light	The bustling energy of ancient commerce fills the marketplace, hopeful anticipation rippling through the crowd as torchlight dances with restless ambition across weathered stone
Male Portrait (portrait)	Historical gentleman portrait, slight head turn, neutral expression, subtle eye movement, formal pose, period lighting	The nobleman's gaze carries quiet authority hardened by duty, a flicker of resolve crossing his weathered features as the weight of responsibility settles behind composed dignity
Burghley House (architecture)	Grand manor house exterior, clouds moving overhead, trees swaying, warm afternoon sunlight, shadows shifting across stone facade	Majestic grandeur radiates from the ancient estate, centuries of ambition and legacy breathing through its towering walls as golden light bestows a reverent glow upon weathered stone
Cherry Blossom (nature)	Japanese path lined with cherry trees, petals falling slowly, light filtering through branches, gentle breeze moving leaves	Ephemeral beauty drifts through the air as cherry blossoms surrender to the gentle breeze, bittersweet impermanence painting the path in soft pink as fleeting joy settles like a whisper

Table 10. Complete STOCK/NEURO prompt pairs for the diverse image generalization benchmark. STOCK prompts describe physical motion; NEURO prompts describe emotional affect grounded in the scene.